(19)日本国特許庁（ＪＰ）　(12) 公表特許公報（Ａ）　(11)特許出願公表番号
特表2001－518665
（P2001－518665A）
(43)公表日　平成13年10月16日(2001.10.16)

(51)Int.Cl.[7]	識別記号	FI		テーマコード(参考)
G06F 15/177	678	G06F	15/177	678C
11/20	310		11/20	310A
13/00	301		13/00	301P
15/173			15/173	Z
H04L 1/22		H04L	1/22	

審査請求　有　　予備審査請求　有　（全 58 頁）　最終頁に続く

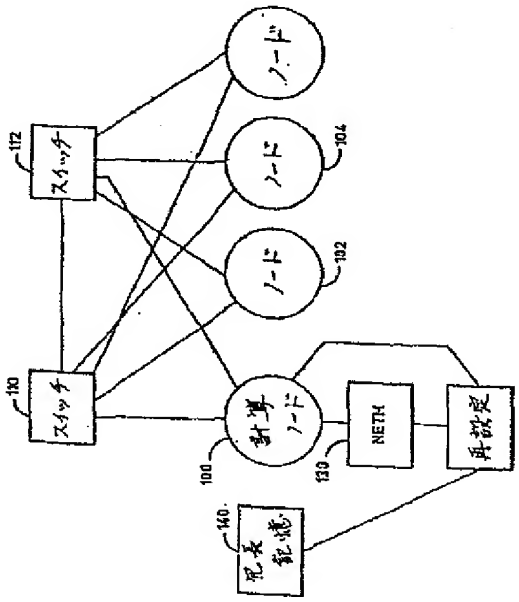
(21)出願番号　特願2000－514214(P2000－514214)
(86)(22)出願日　平成10年10月1日(1998.10.1)
(85)翻訳文提出日　平成12年3月31日(2000.3.31)
(86)国際出願番号　PCT／US98／20532
(87)国際公開番号　WO99／17217
(87)国際公開日　平成11年4月8日(1999.4.8)
(31)優先権主張番号　08／943，049
(32)優先日　平成9年10月1日(1997.10.1)
(33)優先権主張国　米国（US）

(71)出願人　カリフォルニア・インスティテュート・オブ・テクノロジー
CALIFORNIA INSTITUTE OF TECHNOLOGY
アメリカ合衆国91125カリフォルニア州パサディナ、イースト・カリフォルニア・ブールバード1200番
(72)発明者　ブルック，ジョシュア
アメリカ合衆国、91011　カリフォルニア州、ラ・カナダ、ブランブルウッド・ロード、5657
(74)代理人　弁理士　深見　久郎　（外５名）

最終頁に続く

(54)【発明の名称】　分散型計算ノードの高信頼アレイ

(57)【要約】
頑健な分散型サーバシステムを提供するために冗長記憶および冗長通信を利用するシステム。



【特許請求の範囲】

【請求項１】　冗長分散型ネットワークシステムであって、

複数のシステムノードを含み、前記システムノードの各々は少なくとも２つの通信装置および記憶装置を有し、前記記憶装置はネットワークのための情報の冗長記憶を有し、さらに

前記システムノードの前記通信装置に、いずれか１つのシステムノード内の前記通信装置の各々が複数のスイッチング装置の異なる１つに接続されて、前記システムノードの各々が少なくとも２つの異なる経路のうちの１つを介して互いに通信することができ、したがって冗長通信を提供するように接続される、複数のスイッチング装置と、

ネットワークシステムの動作を妨害し得るシステム機能状態を検出する、検出ルーチンと、

ネットワークの冗長性を利用してシステム機能状態を補償するよう、前記通信冗長性または前記記憶冗長性のうち少なくとも１つを使用してネットワークを再設定する、論理ネットワークプロセスとを含む、システム。

【請求項２】　前記検出ルーチンは、前記システムノードの少なくとも複数上のランを検出し、前記システムノードの各々上では各前記検出ルーチンによって同じプロトコルが実行される、請求項１に記載のシステム。

【請求項３】　前記システム機能状態は、ネットワーク通信内の障害、メモリ記憶内の障害、または、望ましくない結果をもたらす他の種類の障害を含む、請求項１に記載のシステム。

【請求項４】　前記システム機能状態は、ネットワーク通信における障害を含み、前記論理ネットワークプロセスは、接続が異なる接続に変更されるよう指令する、請求項３に記載のシステム。

【請求項５】　前記システム機能状態がメモリ記憶における障害を含むときには、前記論理ネットワークプロセスは、所望の情報が前記冗長データ記憶から得られるよう指令する、請求項４に記載のシステム。

【請求項６】　前記検出ルーチンは各ノードで動作して、ネットワーク内の他のノードに対する接続状態を見る、請求項１に記載のシステム。

【請求項７】　前記検出ルーチンは、ネットワークの状態に関するヒントを使用して前記接続状態を判断するよう動作する、請求項６に記載のシステム。

【請求項８】　前記ヒントは、所定の間隔で各前記ノードによって生成されるハートビート信号を含み、前記検出ルーチンは、前記ハートビート信号を受信して前記ヒントの１つとしての前記ハートビート信号の存在または不在を検出するよう動作する、請求項７に記載のシステム。

【請求項９】　トークンパッシングシステムをさらに含み、各ノードは監視されるチャネルを介して監視されるノードにおけるイベントを判断し、前記イベントを示すよう前記監視されるチャネルを介して前記監視されるノードにトークンを渡し、前記監視されるノードは、前記イベントに基づいた動作を示すよう前記トークンを戻し、各ノードは、該ノードとは異なるノード上で対応するイベントが起きずとも該ノードにおいて起こすことのできるイベントの数を制限するよう、所定数のトークンのみを有する、請求項７に記載のシステム。

【請求項１０】　各前記ノードがネットワークの同じ履歴を見ることを保証する手段をさらに含む、請求項６に記載のシステム。

【請求項１１】　前記接続は、計算ノードのグループが分離されることのないように行なわれる、請求項１に記載のシステム。

【請求項１２】　前記スイッチは、可能な限り最も非ローカルな方法で前記ノードを接続する、請求項１に記載のシステム。

【請求項１３】　前記スイッチは、互いに最も遠い２つのノード間を接続する、請求項１２に記載のシステム。

【請求項１４】　前記接続は、どの２つのノードの故障によってもノードの１グループをノードの他のグループとの通信から分離することのないように行なわれる、請求項１３に記載のシステム。

【請求項１５】　各ノードは、少なくとも２つの経路によって各他ノードと接続され、

ネットワークモニタをさらに含み、該ネットワークモニタは、各前記ノードにおいて稼動して、前記ネットワークモニタがその上で稼動するローカルノードから各リモートノードへの各接続経路上のすべての接続を監視する、請求項１に記

載のシステム。

【請求項１６】　高信頼ユーザデータプロトコルをさらに含み、該プロトコルは、前記ローカルノード上で稼動し、該ローカルノードから他のノードへの通信リクエストを受信し、かつ、前記ネットワーク監視プロセスから経路を判断する、請求項１５に記載のシステム。

【請求項１７】　ノード間の物理的な接続を、異なるノード接続に変更することを可能にする、論理ネットワーク相互接続を使用した前記通信経路の再設定をさらに含む、請求項１６に記載のシステム。

【請求項１８】　前記ノード間の動作可能な接続を判断するネットワークモニタと、稼働中のノードのための情報を処理する高信頼ユーザプロトコルと、前記動作可能な接続に基づいて通信を再設定する論理ネットワークと、をさらに含む、請求項１に記載のシステム。

【請求項１９】　前記記憶装置は、各ノードの各ディスク上の情報の一部のみを記憶する、請求項１に記載のシステム。

【請求項２０】　各ノードの各ディスクは、他のディスク上の情報の特性を示す情報もまた記憶する、請求項１９に記載のシステム。

【請求項２１】　冗長分散型サーバであって、

分散型計算ノードのアレイを含み、前記計算ノードの各々は各他ノードとは異なる情報を記憶し、前記記憶される情報は前記計算ノード間で冗長となっており、さらに

前記計算ノードのアレイに接続され、前記計算ノードのアレイ間に冗長通信経路を提供し、どのような所定数のネットワーク障害もシステムの残りのノードの動作に影響を及ぼすことのないように動作する、スイッチングシステムを含み、

前記計算ノードの各々はネットワークステータスを判断する同じプロトコルを実行し、よって、各前記計算ノードが同じネットワーク履歴を見るようにする、サーバ。

【請求項２２】　前記各ノード上に記憶される情報は、所望の情報のすべてではない一部のみを記憶し、どの２つのノードも同じ情報を記憶することはない、請求項２１に記載のサーバ。

【請求項２３】　前記記憶される情報は、情報部分および冗長部分を含み、前記冗長部分は他のノードのための情報部分のみを示す情報である、請求項２２に記載のサーバ。

【請求項２４】　前記冗長部分は、複数の前記ノードが前記情報部分を形成するようアレイへと配列されているアレイ符号から形成され、かつ、前記冗長部分は、前記アレイの対角線方向に沿ったチェックサムで形成される、請求項２３に記載のサーバ。

【請求項２５】　集合的にシステムデータを記憶する複数のノードを含み、各ノードは生のデータおよび、前記各ノードとは異なる他のノード内に記憶されている生のデータを示す冗長データを記憶し、さらに

前記複数のノードからの分散型の読出を実行するよう動作する制御プロセスを含み、前記分散型読出は、ノードの可用性に関するパラメータを判断するステップと、もし前記パラメータが可用性を示す場合には前記複数のノードから前記生のデータを読出すステップと、もし前記パラメータが可用性を示さない場合には前記複数のノードよりも少ない数のノードから前記生のデータおよび前記冗長データの両方を読出すステップとを含む、冗長ネットワーク。

【請求項２６】　誤り訂正符号を示す生の情報および冗長情報を複数の情報ノードに記憶するステップと、

前記情報ノードのユーザビリティを示すパラメータを判断するステップと、

前記パラメータが前記複数のノードが使用可能であることを示す場合には前記複数のノードから前記生の情報を読出し、かつ、前記パラメータが前記複数のノードのうち少なくとも１部分が使用可能ではないことを示す場合には前記複数のノードよりも少ない数のノードから前記生のデータおよび前記冗長データの両方を読出すステップとを含む、誤り訂正符号を使用する方法。

【請求項２７】　ノードからの情報を表わす、アレイの各列を形成することによって、情報のアレイを形成するステップと、

データを示す生の情報を含む、各列の生の部分を形成するステップと、

冗長性に関する情報を示す、冗長情報を形成するステップとを含み、前記冗長情報は、前記各ノード以外の他のノードに関する情報を示し、該情報は、前記他

のノードから情報を得る特定の形状のエンベロープに沿って取られる、ノードのアレイのための符号化方法。

【請求項２８】　前記エンベロープは、前記アレイの縁を超えて他ノードに延びる対角線である、請求項２７に記載の符号化方法。

【請求項２９】　アレイの列に各ノードをマッピングするステップと、

前記アレイの列から冗長情報の２つの行を形成し、前記２つの行を前記列内に位置付けることによって、N－2 ×N個の情報シンボルおよび２×N個の冗長情報シンボルを含む、結果として得られるN×Nのアレイを形成するステップとを含み、前記パリティシンボルは以下の式に従って構築される、複数のノードで形成される冗長符号化ネットワークを形成する方法。

【数１】

$$C_{n-2,i} = \sum_{K=0}^{n-3} C_{k,(i+k+2)_n}$$

$$C_{n-1,i} = \sum_{K=0}^{n-3} C_{k,(i-k-2)_n}$$

式中、$i = 0, 1, \cdots, n-1$ および $(x)_n = X \bmod n$ である。

【請求項３０】　冗長ビデオサーバシステムであって、

複数のシステムビデオ記憶ノードを含み、前記システムビデオ記憶ノードの各々は、少なくとも２つの通信装置および記憶装置を含み、前記記憶装置はビデオ情報の冗長記憶を含み、さらに

前記システムビデオ記憶ノードの前記通信装置に、いずれか１つのシステムノードにおける前記通信装置の各々が複数のスイッチング装置の異なる１つに接続されて、前記システムノードの各々が少なくとも２つの異なる経路のうち１つを介して互いに通信することができ、したがって、冗長通信を提供することができるように接続される、複数のスイッチング装置と、

特定の数の故障がビデオサーバの動作を中断することのないような形をとる、前記スイッチング装置およびビデオ情報の冗長記憶上の接続と、

ネットワークシステムの動作を妨害し得るシステム機能状態を検出する、検出ルーチンと、

ネットワークの冗長性を使用してシステム機能状態を補償するよう、前記通信冗長性および前記記憶冗長性のうち少なくとも１つを使用してネットワークを再設定する、論理ネットワークプロセスとを含む、システム。

【発明の詳細な説明】

【０００１】

【技術分野】

本願は、頑健な通信ならびに分散型の読出および書込動作を形成するような形で、情報の冗長通信および記憶を含むネットワークを形成する、分散型計算ノードの高信頼アレイに関する。また、該システムは、冗長性が必要であることを示す状態を検出し、そのような状態に対応して該状態に対する補償をするために再設定を行なうこともできる。

【０００２】

【背景技術】

分散型環境上の計算および記憶を使用すれば、既存のハードウェアおよびソフトウェアを増強することのできる可能性は大いにある。このようなシステムは、分散型の、可用性に優れた記憶サーバとして使用することができるであろう。可能な用途としては、マルチメディアサーバ、ウェブサーバ、およびデータベースサーバ等がある。しかし、より一般的には、この種のシステムは、情報を複数の場所で分散する必要のある、どのような応用にも使用することができる。

【０００３】

しかしながら、コストを過度に増すことなく高い信頼性を提供するよう、接続、監視および動作を正しく組合せることは、非常に困難である。

【０００４】

障害を補償することのできる冗長記憶システムを提供する方法はよく知られている。このようなシステムの一例が、「ＲＡＩＤ (reliable array of independent disks)」と呼ばれるシステムである。ＲＡＩＤタイプのシステムの２つの例が、米国特許番号第５，５７９，４７５号および第５，４１２，６６１号に記載されている。これらのシステムは冗長データ記憶を提供することで、システムのどのディスクの故障も、該システム内の他の場所にある冗長データによって補償されるようにしている。

【０００５】

システム内の各コンピュータ（「ノード」）が他のノードと接続されている通

信システムはよく知られている。その一例がイーサネットであるが、これはバス型プロトコルであって、計算ノードがバスを介して通信する。サーバは典型的に、すべてのノードのための共用データをすべて格納している。これらのノードは、ローカルなデータ記憶もまた有してもよい。

【０００６】

単一のネットワークシステムは、ノードとサーバとをつなぐ単一のイーサネットリンクを含む。したがって、サーバへの接続もしくは通信に、またはサーバ自身の内部になんらかの障害が発生すると、ノードはもはやサーバから従来のデータアクセスサービスを得ることができなくなる。それらノードは、その後はスタンドアロンモードで動作することを強いられ、したがってローカルに利用可能なデータを使用して動作することしかできなくなる。

【０００７】

このようなシステムの信頼性を高めることを目的としたサーバ型システムが知られている。このようなシステムの１つは、デュアルバス接続を使用する。各計算ノードには、２つの別個のイーサネットカードを使用して２つの別個のバスを介して２つの別個のサーバにつながる、２つのイーサネット接続が提供される。これは、実質的には、各々がハードウェアおよび記憶の完全な予備を有する、２つの別個のシステムである。

【０００８】

この場合、いずれかの接続またはバスにエラーがあっても、他方のバスを介して正常な動作を続けることができる。２つの冗長バスおよび２つの冗長サーバを有するシステムを、デュアルバス、デュアルサーバと呼ぶ。このようなデュアルバス、デュアルサーバシステムは、単一のネットワーク障害は許容することができる。しかし、このようなシステムでは、通常、各サーバ上にすべての情報の複製を作成しなければならない。

【０００９】

【発明の開示】

本願に記載されるシステムは、パーソナルコンピュータ等の比較的ローパワーのワークステーションを使用して、既存のハードウェアおよびソフトウェアを増

強するものである。これらのパーソナルコンピュータは、冗長接続によって接続される。この接続は、既存のハードウェア、たとえばローカルおよび／またはワイドエリアネットワークを利用することができる。

【００１０】

本願は、分散型計算ノードのアレイで形成された、冗長分散型サーバを開示する。それら計算ノードの各々は、特別な冗長方式で情報を記憶し、また、頑健な通信を保証するプロトコルを実行する。

【００１１】

該システムは、好ましくは、ある特定の数のネットワーク障害が、システムの残りのノードの動作に影響を及ぼすことのないように、ネットワークにフォールトトレランスをもたらす、特別なアーキテクチャおよび動作を含む。しかし、そのノードはどれ一つをとっても、すべての情報を二重に記憶することはない。

【００１２】

該サーバシステムは、冗長通信および冗長記憶を含む。冗長通信は、各ノードが少なくとも２つの異なる経路のうちの１つを介して他の各ノードと通信することができるようにする、システムアーキテクチャによって得られる。冗長記憶は、特別な冗長符号化方式を使用して情報を冗長に記憶することによって得られる。

【００１３】

該サーバシステムはまた、システムの機能状態を検出する分散型検出ルーチンを実行する。システムの機能状態としては、たとえば、ネットワーク障害がある。ネットワーク障害は、リンク断、または動作不能ノードもしくは動作不能スイッチングデバイス等の、通信障害を含み得る。しかし、より一般的には、システム機能状態は、ネットワークの動作を妨げるおそれのある、どのような状態をも含み得る。システム機能状態は、システムの冗長性によって補償することができる。

【００１４】

該サーバシステムは、好ましくは、システムの機能状態を検出するネットワーク監視プロセスを実行する。論理ネットワークプロセスがシステムを再設定して

、冗長性を利用してシステム機能状態を補償するようにしている。

【００１５】

該システムはまた、システム障害があった場合に代替的な動作ができるようにする、分散型読出および書込システムを使用する。この代替的な動作は、システムの冗長性を使用する。

【００１６】

【発明を実施するための最良の形態】

本発明の目的、利点および特徴は、添付の図面を参照して以下の詳細な説明を読むことにより、より容易に理解されるであろう。

【００１７】

図１は、高信頼冗長分散型ネットワークサーバシステムの、最も基本的な第１の実施例を示す。このシステムは、複数の計算ノード（「ノード」）と、それらノードの間でスイッチングを行なうネットワークとで形成されている。

【００１８】

図１のネットワークは、ノードとネットワークとの間に、通信および記憶の双方について冗長性を有する。この冗長性は、所定の数のシステム機能状態を補償するのに使用することができる。この冗長性によって補償されるシステム機能状態は、ネットワークにおける障害（「通信障害」）、ディスク、揮発性メモリ、もしくはデータを記憶する他のいかなる種類のメモリであってもよい、メモリの記憶における障害（「メモリ障害」）、または、望ましくない結果をもたらす他のどのような種類の障害をも含み得る。

【００１９】

この分散型サーバシステムはまた、検出プロセスを含む。検出プロセスは、各ノードにおいて稼動して、ネットワーク内の他のノードへの接続状態を見る。各ノードは、ネットワークの状態に関するヒントの集積を使用して、同じプロトコルに従ってネットワークを見る。この検出プロセスは、両サイドがネットワークの同じ履歴を見ることができるようにしている。この検出プロセスは、分散されてはいるが、トークンパッシングシステムを使用して、所望のしきい値内で整合する、ネットワークのノード間のネットワーク履歴を維持する。トークンにより

、一方側が、他方側がアクションをとったことを確認せずにとることのできるアクションの数を、所定数に制限することによって、２つの側の自由度が制限される。

【００２０】

この検出プロセスは、他のプログラムおよびユーザアプリケーションから見えないところで実行される。検出プロセスの好ましい形態は、監視対象となるシステムに関する情報を集めるよう動作する、ネットワーク監視（「ＮＥＴＭ」）プロセスを使用するものである。このＮＥＴＭプロセスは、好ましくは、他方ノードが正しく動作しているかどうかを判断するが、より一般的には、ＮＥＴＭプロセスは、ユーザビリティに関するパラメータを判断する。このパラメータは、以下に説明するように、システムが動作中かダウンしているかを示す表示を含み得る。このパラメータは、システムがどの程度ビジーであるかを示す表示もまた含み得るが、この表示は、負荷の最適配分に使用することができる。

【００２１】

図１のシステムは、２つのスイッチ１１０および１１２によって接続される４つの計算ノード（「ノード」）１００、１０２、１０４および１０６を用いて、本発明の特徴を示している。各ノードは、２つの異なる、したがって冗長な経路を介して、他の各ノードと通信することができる。たとえば、ノード１００は、ノード１００とスイッチ１１０との間の相互接続１２０を介して、ノード１０６と通信することができる。また、スイッチ１１０からノード１０６への経路１２２を介して冗長相互接続を与える、完璧に別個の経路が存在する。これとは別に、ノード１００は、ノード１００からスイッチ１１２への相互接続１２４、および、スイッチ１１２からノード１０６への相互接続１２６を使用して、ノード１０６と通信することもできる。したがって、各ノードは、少なくとも２つの完全に別個の、冗長な接続経路によって、他の各ノードと接続されている。

【００２２】

この冗長通信能力は、ある通信リンクの使用を避ける方が好ましい場合に、異なる経路を選択することができるようする。たとえば、スイッチ１１０、または、接続線１２０および／または１２２のどこかの部分が失われた場合にも、スイ

ッチ１１２を介して、接続線１２４および１２６を通じて、通信を行なうことができる。

【００２３】

情報もまた、冗長な方法で記憶されており、ネットワークのどこかの部分が故障するか、他の形、たとえば膨大なトラフィック等により利用不可能となった場合にも、どのような情報も検索することができる。この冗長記憶機構を、図１においては要素１４０として示す。システム内のノードの全個数をｎで表わし、選択数をｋで表わすとすると、冗長記憶１４０内のデータは、ｎ－ｋ個のノードが失われた場合にも、システムから所望のデータを得ることのできる能力に影響が及ぶことのないような態様で、記憶されるのが好ましい。このような記憶は、好ましくは、ノードの各々に冗長情報が記憶される最大距離分離（「ＭＤＳ」）符号方式に従ってデータを記憶することによって行なわれる。この冗長情報を他のノードのデータとともに使用して、欠けてしまった１つ以上のノードのためのデータを再構築することができる。

【００２４】

検出プロセスによって、動作不能ノードや破断した通信リンク等の望ましくないシステム機能状態が検出されると、再設定プロセス１４０が実行される。この再設定プロセス１４０は、記憶冗長性または通信冗長性の少なくとも一方を使用することができるというその能力のおかげで、障害に対して頑健である。再設定プロセスは、特定の障害があった場合にもシステムが動作することができるようにする。しかし、このプロセスは、専用のスイッチングを何ら必要としないであろう。たとえば、ノード１００と１０６との間の経路は、１２０／１１０／１２２を介する経路１上で、または、１２４／１１２／１２６を介する経路２上で、構築することができる。正常動作中には、通信は、経路１、次に経路２、その次に経路１、等のように、交互に行なわれる。しかし、経路１内に障害または過負荷があれば、すべての通信は経路２上で行なわれるようになる。これは、通信が適切に方向付けられるという意味で、再設定と言える。いずれにせよ通信の半分が経路２を介するように方向付けられていたであろうとはいうものの、この再設定によって、すべての通信が経路２を介するように仕向けられるのである。

【００２５】

したがって、図１は、本願において説明する分散型サーバの基本的な特徴を示している。これらの特徴は、通信の冗長性、記憶の冗長性、その冗長性によって補償することのできるイベントの検出、および、そのイベントを補償するために冗長性を使用する再設定、を含む。

【００２６】

冗長通信

図１のシステムは、４つのノード１００～１０６ならびに２つのスイッチ１１０および１１２を含む、簡単な冗長接続を示す。これらのノードは、好ましくは、各々が２つのＰＣＩバス型通信カードを有する、パーソナルコンピュータ（「ＰＣＳ」）等のスタンドアロンワークステーションである。通信カードは、スイッチを介して、他のＰＣＳの同様の通信カードと通信する。通信カードのプロトコルは、イーサネットまたは他の市販のどのような種類のものであってもよい。好ましいシステムは、図２に示すスイッチングノード２００に対して、ミリネット（Ｍｙｒｉｎｅｔ）スイッチを使用する。ミリネットスイッチは市販されており、また、ボーデン（Boden）他による「ミリネット：ギガビット／秒のローカルエリアネットワーク（"Myrinet: a gigabit per second local area network"）」IEEE Micro 1995、に記載されている。

【００２７】

本発明によって使用される特別なノード接続は、ネットワーク通信障害があった場合にも正常に動作する能力を改善する、通信冗長性を提供する。ネットワーク通信障害は、スイッチ障害、リンク断、またはスイッチ故障等の、障害を起こした通信を含む。接続は、単一の通信障害または通信障害の組合せが通信の破壊またはノードの孤立化を引起す可能性を最小に抑えるような方法で構築される。正しい接続の重要性を、以下に説明する。

【００２８】

図２は、４つのスイッチ２２０～２２６を使用して８つの計算ノード２００～２１４を接続するシステムを示す。各計算ノードは、使用可能な２つの相互接続リンク経路を含み、これにより、通信の冗長性が与えられている。

【００２９】

しかし、図２のシステムにおける通信故障は、計算ノードのグループを意に反して「分離する」おそれがある。これらの孤立した計算ノードのグループは、それらが分散型サーバの他の稼動中のノードのすべてともはや通信することができないという意味で、分離されている。

【００３０】

１例として、もしスイッチ２２４および２２６が両方とも故障してしまえば、計算ノード２００～２０６が計算ノード２０８～２１４とは完全に分離してしまう。この場合、使用可能ではあるがあまり好ましくない状態の、分離し得るシステムができることになる。

【００３１】

たとえば、使用されるＭＤＳ符号が、データを再構築するのに６個から８個のノードを必要とするものとする。もしこのシステムが上述のように分離されれば、半数のノードのみしか通信ができなくなる。通信可能なノードは４つなので、この場合は障害によって、データの再構築ができなくなってしまう。

【００３２】

図３のような接続構造が好まれる。この１０ノード、４スイッチを含むシステムでは、通信障害がある場合の相互接続が改善されている。コネクションインターフェイスは、いずれか２つのスイッチが失われても、最悪の場合でも２つの計算ノードのみにしか影響が及ぶことのないように作られている。たとえば、スイッチ３２０および３２６が故障した状態を示す図４を参照されたい。太い線は、この故障によって影響を受ける通信線を示す。この２つのスイッチの故障によって分離されるのは、計算ノード３０４および３１２のみである。他のすべてのノードは、全く問題なく動作でき、どれも分離されることはない。

【００３３】

フォールトトレランスの重要な部分は、スイッチおよびノードの特定的な相互接続によって得られる。上述の例のように、図２のシステムは、計算ノードの半分ずつが分離されてしまうおそれがある、という欠点を有する。分離されたシステムは計算ノード２００～２０６を含み、それらのノードは通信することはでき

るが、ノード２０８～２１４のグループとは分離されている。

【００３４】

この問題の別の例を、図１２に示す。図１２は、複数のスイッチングノードを使用して、多数の計算ノードを相互接続する方法の１例を示す。各スイッチングノードＮは、隣接する２つの計算ノードＣの間に位置付けられる。これは、使用可能ではあるが、あまり好ましくない構成である。計算ノード１２００および１２０２が同時に障害を起こしてしまった場合、このシステムの通信能力は、図１２に示した破線に沿って分割されてしまうことに留意されたい。これにより、システムの半分１２０４が、システムの他方の半分１２０６とは実質的に分離されてしまう。

【００３５】

本願に開示した接続の目的は、２つの通信故障によって形成されるおそれのある、この種の分離を防ぐことである。好ましいシステムは、可能な限り最も非ローカルな方法でノードを接続する方法を説明する。これは、各スイッチングノードが最も近い２つの計算ノードに接続される図１２のシステムとは、対称的である。発明者は、非ローカルなスイッチ間を接続するこの非自明のシステムが、非常に優れたフォールトトレランスをもたらすことを発見した。

【００３６】

図１３はそのような装置を示す。各ノードは２つのスイッチに接続されており、この図は、最も遠い２つのスイッチ間が接続されている様子を示す。図１３のようにスイッチおよびノードを並べて図示した場合、互いに対して物理的に最も遠い２つのスイッチ間を結ぶ接続は、直径として表わされる。このような接続は、どれか３つのスイッチが失われた場合にも、ユニット全体の半分ずつを分離させるような影響が生ずることはない、という利点を有する。逆に、いずれか２つの場所でユニットに破損が生じた場合にも、多くのノードの間で通信は依然として行なうことができる。どこか３つの場所で損失があった場合でも、システム内のノードの全個数に関係なく、ある一定数のノード－直接影響を受けるノード－のみが分離されるにすぎない。

【００３７】

たとえば、ロケーション１３１０で通信故障が生じ、ロケーション１３１２で別の破損が生じたとする。スイッチ１３００がスイッチ１３０４を介して依然としてスイッチ１３０２に接続しているので、ノードが依然として通信することができるのは明らかである。スイッチ１３００はまた、スイッチ１３０８を介してスイッチ１３０６にも接続されている。同様に、これらのスイッチはすべて、破損があった場合にも互いに接続されている。さらに、この好ましいシステムにおいては、必要とされるであろうノード間接続は、最大でも、システム全周の１／４である。

【００３８】

この非ローカルという概念は、リング型以外の構成にもあてはまる。たとえば、リングと見なすことのできる構成であれば、どのようなものも代替的に使用することができる。

【００３９】

図１から図３に示した好ましいサーバシステムは、ミリネット相互接続技術を使用して冗長ネットワークを介して接続される、パーソナルコンピュータ型のワークステーションを使用する。もちろん、代替的に、１００ＭＢイーサネット等の他の通信技術を使用することもできる。これらのシステムはすべて、共通に、障害を起こしたリンクがあった場合にも冗長性を維持するという能力を有する。このシステムは、どのような数の通信要素とも、ともに使用することができる。ただし、好ましい数、また開示される数は、２である。

【００４０】

冗長記憶

図１の好ましい実施例において、各ノードは、所与の記憶データのうち、１部分のみを記憶している。この記憶データは、ローカルノード内に実際に記憶されている各情報の一部、および他のノードが記憶している一部を使用して、取出される。この概念を図５を参照して説明する。図５は、ビデオサーバを示す。開示されるサーバは、図示されるように表示されるビデオを示すデータを提供する。各計算ノードは、図示するように、データ全体の半分の量を記憶している。このデータは、そのデータを要求する１ノード内に記憶されているデータを、他のノ

ード内に記憶されているデータと組合せることで、どのビデオフレームも再構築することができるような方法で、冗長に記憶されている。

【００４１】

この記憶方式によれば、どのノードも、そのノードが他のすべてのノードから分離されていない限り、そのノードが所望する情報を受取ることができる。この方式は、分散型サーバ内に多数の故障が生じた場合に備える、記憶冗長性を提供する。

【００４２】

しかし、より一般的には、ここに規定する好ましい方式は、全部でｎ個のノードのうち、ｋ個の稼働中のノードのサブセットから、データを再構築することができるようにするものである。ｋ＝２およびｎ＝４の場合の例を、以下に示す。

【００４３】

図６は、計算ノードの１つに故障が生じた場合に、送り出されたビデオのどの項目でも、残りの計算ノードで復元することができる方法を示す。これは、冗長記憶を可能にするどのような符号方式でも達成することができる。

【００４４】

好ましいシステムは、いずれか２つの通信リンクを失っても、そのサーバシステムの他のいかなる通信機能も失うことがなく、かつ、障害を実際に含むノード以外の他ノードに影響を及ぼすことがない。

【００４５】

このシステムの冗長記憶特徴は、各ノード内に、データ全体よりも小さいサイズ、全データの半分の量の、符号化されたデータを記憶する。したがって、ｋ個の稼働中のノードを有するシステムにおけるサイズＫの各ファイルについて、この好ましい実施例においては、そのファイルのＫ／ｋが、サーバの各ノード上に記憶されている。ファイルの他の（ｋ－１）は、他のｋ－１個の稼働中のノードから得られる。

【００４６】

Ｘ符号

記憶の冗長性は、好ましい実施例に従えば、ノード間で情報の記憶を分散させ

ることによって得られる。上述のように、サイズＫの情報の各項目について、好ましいシステムは、各ノード内にＫ／ｋのデータ（情報のオリジナルサイズ）を記憶する。ここで、ｋは、データを再構築するのに必要となるであろうノードの数を表わす。各ノードは、他のいずれかのノードに記憶される情報の他のＫ／ｋにアクセスすることによって、情報のいずれの項目も再構築することができる。情報は、好ましくは、その情報を記憶するために最大距離分離（「ＭＤＳ」）符号を使用して記憶される。情報を記憶するための好ましい形態は、Ｘ符号と呼ばれる新しい符号化システムを使用する。ここに記載するＸ符号は、ノード間、より特定的にはノードのディスク間に分散される、情報の各項目を記憶するための、特別な、最適化された符号である。

【００４７】

最も好ましくは、情報の一部のみ、符号化されたデータのある部分のみが、各ノード上に記憶される。各ノードは、他のノード上の情報のなんらかの特性を示す情報もまた記憶する。この特性とは、たとえば、他のノード上のデータの合計を示す、チェックサムまたはパリティであり得る。この情報は、他のノード上の情報を再構築するために、それら他のノード上の情報とともに使用される。

【００４８】

上述のように、使用される好ましい符号はＸ符号であって、これを以下に詳細に説明する。Ｘ符号は、Ｎ×Ｎの最大距離分離（「ＭＤＳ」）アレイ符号であって、Ｎは好ましくは素数である。この符号は、排他的論理和（「ＸＯＲ」）および循環シフト演算のみを使用して、符号化および復号化することができる。このため、Ｘ符号は、リードソロモン符号等の計算がより複雑な符号に比べて、はるかに高速で符号化および復号化することができる。

【００４９】

このＸ符号は、最小列距離３を有する。これは、この符号が１つの列誤りまたは２つの列消失を訂正することができることを意味する。Ｘ符号は、Ｘ符号内の単一の情報ビットまたはシンボル等の、単一の情報ユニットの変化が、常に、２つのパリティビットまたはシンボルだけにしか影響を及ぼさない、という、特定的な特性を有する。したがって、情報が更新されるたびに、変更が必要となるの

は、それら２つのパリティビットまたはシンボルのみである。

【００５０】

Ｘ符号のシステムは、図１５に示すアレイを使用する。各列１５００は、単一ノード内の情報を表わし、各ノードにマップされる。パリティシンボルは、列ではなく行に記憶される。

【００５１】

この符号は、ネットワークのすべてのノードを使用して配列されて、集合的にＮ×ｎのアレイを形成する。ここで、好ましくはＮ＝ｎである。このアレイは、Ｎ－２ ×Ｎの情報シンボル、および２×ｎのパリティシンボルを含む。図１４（Ａ）は、ｎ＝５の場合のアレイの例を示す。ノード１４００の部分は情報を表わし、四角で囲った要素は、１ビット、１セクタまたはディスクの他の何らかのユニット等の情報の１ユニットを表わす。本明細書では、これらのユニットを総称的にシンボルと呼ぶ。

【００５２】

非情報部分１４０２は、冗長情報を表わす。ここで説明するように、アレイの単一列によって表わされるディスク、たとえばディスク番号１４０４のディスクに対して、冗長情報１４０２は、他のディスクからの冗長情報を表わす。すなわち、この冗長情報は、１４０４以外のディスクからのみ得られるものである。

【００５３】

このＸ符号システムは、ディスク１４０４全体の内容を表わす列を形成する。Ｘ符号のパリティシンボルは、ディスク上の２つの付加的な行１４０２から形成されている。したがって、各ディスクは、Ｎ－２個の情報シンボル、および２個のパリティシンボルを有する。１列内の１シンボルの誤りまたは消失は、列消失から復元することができる。

【００５４】

ここで具体的な符号化手順を説明する。ｉ番目の行およびｊ番目の列のシンボルをC_{ij}とすると、Ｘ符号のパリティシンボルは、以下の式に従って構築される。式１：

【００５５】

【数２】

$$C_{n-2,i} = \sum_{K=0}^{n-3} C_{k,(i+k+2)_n}$$

$$C_{n-1,i} = \sum_{K=0}^{n-3} C_{k,(i-k-2)_n}$$

式中、ｉ＝０，１，…，ｎ－１ および $\langle x \rangle_n = X \bmod n$ である。

【００５６】

これは、幾何学的には、傾き１および－１の対角線方向にそれぞれ沿ったチェックサムを表わす、パリティ行と解釈することができる。

【００５７】

図１４（Ａ）は、第２のパリティチェック行１４１２が存在しないかまたはすべてゼロ（０）であると仮定して、第１のパリティチェック行１４１０を得る方法を示す。なお、存在しないかまたはすべてゼロのパリティチェック行は、仮想ゼロ行と呼ぶ。チェックサムは、傾き－１のすべての対角線方向上で形成される。図１４（Ａ）においては、三角形をすべて加算することで、第１のパリティチェック行１４１０が形成される。具体的には、要素１４１４、１４１６、１４１８および１４２０を加算して、パリティ要素１４２２を形成するのである。

【００５８】

図１４（Ｂ）は、例示の複数の単一ビットについて、第１のパリティチェック行を計算する１例を示す。対角線方向の要素１４１４、１４１６、１４１８および１４２０については、１＋１＋１＋０を計算することでパリティ１が導出され、これがシンボル１４２２として記憶される。

【００５９】

対角線は、アレイの外縁に達すると隣接する行に続く。たとえば、要素１４３２、１４３４、１４３６および１４３８を含む対角線方向の行１４３０は、次の行の最上部に続いて、１４４０で始まる。パリティシンボル１４３６は、シンボル１４３２、１４３４、１４３８および１４４０を加算したものに対応する。図１４（Ｂ）は、これらのシンボルが０＋０＋０＋１に対応して、１となることを示しており、シンボル１４３６として値１が記憶される。

【００６０】

第２のパリティチェック行は、傾き＋１の対角線方向から形成される。図１４（Ｃ）は、図１４（Ｄ）とともに、第２のパリティ行の同様の計算方法の具体例を示す。行１４４０は、シンボル１４４２、１４４４、１４４６、１４４８および１４５０を含む。パリティシンボル１４５０は、１４４２＋１４４４＋１４４８＋１４４６で計算される。図１４（Ｄ）は、具体的な例を示しており、ここでは、＋０＋０＋１＝１から、パリティ０が得られる。

【００６１】

図１４（Ｅ）は、これら２つのパリティチェック行を組合せることによって形成される、完全な符号語を示す。２つのパリティチェック行は、互いに完全に独立して得られる。各情報シンボルは、各パリティ行に厳密に１度だけ現れる。すべてのパリティシンボルは、他の列（他のディスク）からの情報シンボルにのみ依存するものであって、互いに依存しあうものではない。したがって、１つの情報シンボルの更新は、２つのパリティシンボルのみを更新させる結果となる。

【００６２】

上述のＸ符号は、真に対角線方向の計算を考慮して、素数ｎを使用している。しかし、ｎが素数でない場合にも、計算には別の線を使用できる。たとえば、ｎ－１個のディスクのすべてを横断するような好適な所与のエンベロープも、Ｘ符号に従って使用することができる。これらの線はすべて、平行であることが好ましい。

【００６３】

上述のように、Ｘ符号は列距離３を有し、２つの列消失または１つの列誤りを訂正することができるようにしている。消失とは、問題があって、どの領域に問題があるのかがわかっている場合である。誤りは、問題の具体的な原因がわからない場合に生じる。復号化動作は、有限体演算を必要とせず、循環シフトおよび排他的ＯＲのみを使用して、利用することができる。

【００６４】

１つの消失の訂正は、どちらかのパリティ行を使用して傾き１または－１の対角線方向に沿って、その消失を簡単に復元することができる。

【００６５】

サイズＮ×ｎのアレイにおいて、２つの列が「消失」であるものと仮定する。この場合、これら２つの列の基本の未知のシンボルは、それらの列内にある情報シンボルである。各列が（ｎ－２）の情報シンボルを有するので、未知のシンボルの数は２×（ｎ－２）となる。同様に、残りのアレイは、これら２×（ｎ－２）の未知のシンボルを全て含む、２× n−2のパリティシンボルを含む。このため、消失の訂正は、２×（ｎ－２）の線形方程式から２×（ｎ－２）の未知数を算出する問題となる。これらの線形方程式は、線形独立であるため、解くことが可能となる。

【００６６】

さらに、この符号の同じ列内の２つの情報シンボルが、同じパリティシンボル内に現れることはない。したがって、各方程式は、２個以下の未知のシンボルを有することになり、未知のシンボルを１つしか含まない方程式もある。このことは、方程式を解く複雑さを大いに減じる。このシステムに従って使用されるシステムは、１つの知られている未知のシンボルを有する方程式から始める。これらの方程式を解くことは比較的簡単である。このプロセスは、すべての方程式が解けるまで、他の未知の解を求めて続けられる。

【００６７】

消失列がｉ番目およびｊ番目（$0 \leq i < j \leq n-1$）の列であると仮定する。各対角線は、ｎ－１本の列のみを横断するので、１本の対角線が最終行において１本の列と交差する場合には、その列のどのシンボルも、この対角線内に含まれることはない。これにより、２つの消失列の１つのシンボルしか含まないパリティシンボルの位置が決定される。このシンボルは、この対角線に沿った簡単なチェックサムから復元することができる。

【００６８】

まず、傾き１の対角線について考える。ｉ番目の列のｘ番目のシンボルが、１対角線内の唯一の未知のシンボルであると仮定する。この場合、この対角線は（ｎ－１）番目の行においてｊ番目の列に当たり、ｙ番目の列で第１のパリティ行に当たることになる。すなわち、３つの点（ｘ，ｉ）、（ｎ－１，ｊ）および（

ｎ－２，ｙ）は、傾き１の同じ対角線上にあり、したがって、以下の等式が成り立つ。

【００６９】

【数３】

$$\begin{cases}(n-1)-x\equiv j-i \mod n\\(n-1)-x\equiv j-i \mod n\end{cases}$$

$$(n-1)-(n-2)\equiv j-y \mod n$$

式中、$1\le j-i\le n-1$ および $0\le j-1\le n-1$
であるため、ｘおよびｙの解は以下となる。

$$x=<(n-1)-(j-i)>_n=(n-1)-(j-i)$$
$$y=<j-1>_n=j-1$$

【００７０】

したがって、パリティシンボル$C_{n-2,j-1}$は、ｉ番目の列内のシンボル$C_{(n-1)-(j-1),i}$の計算を可能にする。同様に、ｊ番目の列内のシンボル$C_{(j-1)-1,j}$は、パリティシンボル$C_{n-2,<i-1>n}$から直接解くことができる。

【００７１】

傾き－１の対角線方向と対称に、ｉ番目の列内のシンボル$C_{(j-1)-i,i}$はパリティシンボル$C_{n-1,<j+1>n}$から解くことができ、また、ｊ番目の列内のシンボル$C_{(n-1)-(j-1),j}$は、パリティシンボル$C_{n-1,i+1}$から解くことができる。

【００７２】

１つの情報シンボルと、傾き１および－１の対角線とがそれぞれ厳密に１度だけ交差することに留意されたい。未知のシンボルが傾き１（または－１）の対角線に沿って解かれ、その後、その解かれたシンボルと交差する傾き－１（または１）の対角線に沿ってパリティシンボルが解かれると、他の列における別の未知のシンボルを解くことができる。この手順は、パリティシンボルが「消失」列となるか、または、解かれたシンボル自体がパリティシンボルとなるまで、反復的に用いることができる。これら同様の技術を使用して、１つ以上のどのような所

望の未知のシンボルも、復元することができる。

【００７３】

好ましいシステムは、Ｎ＝ｎまたはＮとして素数を使用する。図５および図６のようなシステム（ｎ＝４；ｋ＝２）を上述のように使用することもできる。

【００７４】

分散型読出／書込

本システムは、分散型読出および書込システムを使用することによって、新しい種類の動作を行なうことができる。

【００７５】

情報の冗長記憶により、システムは、システムの帯域幅を最大にするように、ｎ個のノードすべてから読出しを行なうことができる。この場合、システムは、ノードの生の情報部分１５０２からのみ読出しを行なう。

【００７６】

代替的に、ノードのうちｋ個のみを読出すこともできるが、これらｋは、それらのパリティ部分１５０４とともに読出される。従来の「訂正」とは異なり、このシステムは、利用可能なクラスタのうちどれを使用するかを、システムが見たネットワークの状態に基づいて選択する。異なる部分を、たとえば奇偶符号等の異なる符号のために使用することができる。

【００７７】

分散型の書込は、情報が変化するたびに、影響を受けるすべてのノードに書込みを行なう。しかし、更新はできる限り小規模とされる。ＭＤＳ符号は冗長性を保証し、更新を最適に最小かつ効率的にする。平均単位パリティ更新数は、符号内で単一の情報ビットが変化した場合に影響を受ける、パリティビットの平均数を表わす。このパラメータは、記憶アプリケーションにアレイ符号が使用されているときに特に重要となる。Ｘ符号は、各単一の情報ビットの変化が２つのパリティビットの更新しか必要としないという意味で、最適である。

【００７８】

Ｘ符号の別の重要な特徴は、独立なパリティビットが形成されることに起因する。これまで使用されてきた符号の多くは、３という符号距離を形成するために

従属パリティ列に依存している。これらのパリティの計算は、パリティが互いに依存しあっているために、非常に複雑であり得る。このことはしばしば、その符号の平均単位パリティ更新数がアレイの列の数に応じて線形に増加する、という状況を生み出す。

【００７９】

米国特許番号第5,５７９，４７５号に記載されたＥＶＥＮＯＤＤ（奇偶）符号または他の同様のシステムは、独立パリティ列を使用して、情報の更新をより効率的なものにしている。

【００８０】

検出

分散型データ記憶システムは、複数ノードにわたってサーバ機能を広げる。これは、本システムに従えば、多数ノードの各々の上で稼動する、特別な通信層を使用して行なわれる。この通信層は、アプリケーションに対して透過的である。特別な分散型読出システムおよび分散型書込システムは、本システムの頑健な動作もまた維持する。

【００８１】

好ましいシステムの通信アーキテクチャを図７に示す。実際の通信およびネットワークインターフェイスは、要素７００として示されている。通信は、イーサネット、ミリネット、ＡＴＭサーバネット（ATM Servernet）、または従来の他のどのような通信方式をも含む、従来の方法で行なうことができる。これらの従来のネットワークインターフェイスは、冗長通信層によって制御される。

【００８２】

通信は、ネット監視（「ＮＥＴＭ」）プロトコルシステム７０２によって監視される。ＮＥＴＭは、各ノードにおいてチャネル状態およびチャネル状態の履歴を判断する接続プロトコルを維持する。より特定的には、ＮＥＴＭは、ＮＥＴＭがその上で稼動しているローカルノードから各リモートノードへの、そのローカルノードとリモートノードとを結ぶ各接続経路を介した接続のすべてを監視する。ＮＥＴＭは接続チャートを保持するが、このチャートは、ローカルノードから各リモートノードへの可能なすべての接続のステータスを常時知らせる表示を含

む。

【００８３】

実際の通信は、高信頼ユーザデータプロトコル（「ＲＵＤＰ」）によって制御される。ＲＵＤＰは、ローカルノード（「ノードＡ」）が他のノード（「ノードＢ」）への通信を求めるリクエストに基づいて動作する。ＲＵＤＰは、ＮＥＴＭから、ノードＡからノードＢへの正常に動作している通信経路に関する接続情報を入手する。ＲＵＤＰは、ＮＥＴＭによって集められた情報を使用して通信経路を選択し、その情報を束ねられたインターフェイスを使用して送出する。ＲＵＤＰはまた、よく知られているプロトコルシステムを使用して情報を適宜パッケージして、順序正しく確認された配信を行なう。

【００８４】

ＮＥＴＭシステムは、システムの各ノード上で稼動して、システムに関する情報を見つける。ＮＥＴＭは、自身がその上で稼動しているノードを、ローカルノードとして見ている。ＮＥＴＭはシステムクルーを使用して、そのローカルノードとシステム内の他のすべてのノードとの間の接続の状態を判断する。

【００８５】

すべてのノード上で同じプロトコルが稼動しているので、各ノード上の各ＮＥＴＭプロセスは、ＡからＢへのどの接続状態についても、同じ条件を判断するであろう。ＮＥＴＭはまた、履歴チェッキング機構を使用して、すべてのノードがある時間にわたってチャネル状態の同じ履歴を見るようにしている。

【００８６】

好ましいシステムクルーは、ノードＡからシステム内の他の各ノードへと、そのノードへの可能な各経路上を送られるメッセージから得られる。これらのメッセージは「ハートビート」と呼ばれる。ＮＥＴＭは、ローカルノード（「ノードＡ」）から各リモートノード（「ノードＢ」）へと、各経路上でメッセージを送る。各接続は、ｉ＝ローカルインターフェイス、ｊ＝リモートノード、およびｋ＝リモートインターフェイスを含む、Ｃｉ，ｊ，ｋ「タプル」と呼ばれる情報の３つの項目によって特徴付けられる。このタプルは、明確な経路を規定する。

【００８７】

ＮＥＴＭは、ハートビートを使用して、各経路Ｃｉ，ｊ，ｋ上にＡとＢとの間に運用中の通信リンクが存在するかどうかを判断する。ＮＥＴＭプロトコルはノードＢ上でも稼動しているので、そのリモートＮＥＴＭも、各経路Ｃｉ，ｊ，ｋを介したノードＢからＡへの接続状態について、同じ判断を行なうであろう。

【００８８】

たとえばバッファオーバフロー等のような障害は、一方方向のみでチャネルが失われる原因となり得る。接続プロトコルはトークンパッシングシステムを使用して、チャネルの履歴を対称にする。

【００８９】

履歴の検出は、接続の動作可能性に関するヒントの集積に基づく。ハートビートは好ましいヒントであり、ここにさらに詳細に説明する。別のヒントの例としては、通信ハードウェア、たとえばミリネットカードからの、障害表示がある。経路Ｘ上の通信を制御しているミリネットカードが、動作不能であるとの表示をすると、プロトコルはその経路が動作不能であると評価することができる。

【００９０】

ヒントの集積は、Ｘを介したＡからＢへの通信経路の状態を評価する変数の状態を設定するのに使用される。この変数は、稼働中（アップ）に対してＵ、ダウンに対してＤの値を有する。

【００９１】

この動作を、図８の概略的なフローチャートに示す。図８の実施例は、信頼性の低いメッセージで形成されたハートビートメッセージを使用している。高信頼メッセージ通信システムでは、送信側のノードがメッセージの受信確認を受取ることが要求される。送信側ノードは、何らかのメッセージ受信確認を受取るまで、そのメッセージを送り続けるのである。これに対し、図８のシステムは、信頼性の低いメッセージ通信を採用している。すなわち、メッセージが単に送信されるのみであり、受信の確認が受取られることはない。

【００９２】

メッセージ８００は、低信頼パッケージメッセージとしてノードＢに送られる。ハートビートは、好ましくは１０ミリセカンド（ｍｓ）毎に送られる。システ

ムは、ステップ８０２においてネットワークヒントを待ってチェックして、ネットワークリンクの状態および履歴を評価する。ハートビートは、あるノードから他ノードに送信されるどのようなメッセージであってもよい。

【００９３】

各ノード上では同じプロトコルが稼動しているので、各ノードは、１０ｍｓ毎に他の各ノードからハートビートを受信すべきであることをわかっている。各ＮＥＴＭにおいて、そのＮＥＴＭが他のノードからハートビートを受信するたびにリセットされるタイマが稼動している。他のノードからハートビートを受信する前にタイマが時間切れとなった場合、その接続に何か問題がある、と判定することができる。

【００９４】

各サイドは、ある時間にわたって同じ履歴を見ていることを確認しようとする。これは、ポイント・ツー・ポイントプロトコルを構成しているノードの対の間で高信頼トークンを渡すことによって行なわれる。各トークンは、ノードがイベントを見たことを示す。そのトークンが他のノードによって受取られる場合、そのノードもまた、対応するイベントを見てトークンを送ったことを意味する。各サイドは、イベントを認めるとトークンを渡す。これにより、両サイドにおける履歴を同じものに維持することができる。

【００９５】

各サイドが渡すことのできるトークンの数は有限である。これは、他のノードによってイベントが確認される前に起こり得るイベントの数を制限する効果がある。たとえば、１サイド当り２つのトークンが当初あったとすると、ノードは渡せるトークンを２つしか持っていないことになる。チャネルの状態に変化が認められるたびにトークンが１つ渡される。他サイドからトークンが届かなければ、２つの変化を認めた後には、そのノードの所有するトークンはなくなる。このことは、各ノードが、他ノードよりも２つのイベントまたはアクションのみしか先行する（または遅れる）ことができないことを意味する。このトークンの手渡しは、２つのノード間の自由度を制限する。すなわち、１つのノードがチャネルがダウン状態であるという報告を受けてから、他ノード側がその情報を得るのを待

つ間、それら２つのノードが離れて動作することのできる度合いを制限する。

【００９６】

別の見方をすれば、それらトークンは、１つのノードが他ノードが同様にアクションを取ったことを知る前に行なうことのできる、遷移の最大数を設定する。

【００９７】

ＮＥＴＭシステムの好ましい実施例を、図９の接続プロトコルステートマシンにおいて示す。図１０Ａおよび図１０Ｂは、その実施例のフローチャートである。ステップ１０００は、１つのノードから他の知られているすべてのノードへの物理的に可能なすべてのチャネルの各々について、ローカルインターフェイスＩＤ、リモートマシンＩＤ、およびリモートインターフェイスＩＤを含むＣｉ，ｊ，ｋの３タプルを形成する最初のステップを含む。プロセスＣｏｎｎＰ（Ｃｉ，ｊ，ｋ）は、すべてのＣｉ，ｊ，ｋの３タプルに対して実行されて、これらチャネルの各々について接続状態が判断される。これにより、各C_iチャネルについてアップ／ダウン（１または０）のステータスを示すブール値を記憶する、Ｃｏｎｎｅｃｔｅｄ（C_i，ｊ，ｋ）と呼ばれるデータ構造が形成される。

【００９８】

ステップ１００２は、ＣｏｎｎＰ（C_i，ｊ，ｋ）イベントがあったかどうかを判断する。もしイベントがなかった場合、実行すべきことは何もないので、プロセスは初めに戻る。

【００９９】

ステップ１００２においてイベントが検出されれば、フローはステップ１００４に進み、そこで、そのイベントがシステムアップイベントであるかどうかが判定される。そうであれば、結果として「１」が戻され、そうでなければ、結果として「０」が戻される。

【０１００】

図１０Ｂのリンクステータスのフローチャートは、他のエンドポイントシステムの動作の証拠として「トークン」のカウントを使用する。

【０１０１】

ステップ１０１０で、プロセスは、最初の値ｎ≧２に設定されているトークン

カウント（「ｔ」）で開始する。システムは、ステップ１０１２で、アップ（「１」）に最初に設定されている状態で開始する。ステップ１０１４は、タイム・インイベントがあったかどうかを検出する。タイム・インイベントは、たとえば、ノードＢからハートビートを受信することによって引起される。状態はこの時点で既にアップなので、タイム・インイベントが検出されても状態はアップのままであって、さらなるアクションが取られることはない。ステップ１０１４においてタイム・インイベントがなかった場合には、ステップ１０１６において、たとえばタイマが時間切れとなる前に予測されるハートビートが受取られなかったために生じる、タイム・アウトイベントの判断が行なわれ、ステップ１０１８においては、トークンが受信されたかどうか（「トークン到着イベント」）が判断される。それらのイベントが起こっていなかった場合には、制御は再びステップ１０１２に渡され、それらのイベントが起きないかどうかの監視が続けられる。システムはこの時点では必ずトークンを持っているため、別のトークンを調べる必要はない。

【０１０２】

ステップ１０１６においてタイム・アウトイベントがあるということは、経路Ｘを介してノードＢからハートビートが受信されなかったことを意味し、したがって、経路Ｘを介したノードＢへの通信に問題がある可能性を意味する。そこで、制御がステップ１０２０に移る。ステップ１０２０においては、タイム・アウトイベントを示すトークンがノードＢに送られて、所定の時間内にハートビートが届かなかったことが報告される。トークンが送出されたので、トークンのカウントもまた１０２０においてデクリメントされる。この後、ステップ１０２２においてＣｏｎｎＰの状態がＤに変更される。

【０１０３】

ステップ１０１８においてトークン到着イベントがあると、１０２４においてトークンの受信ステップが行なわれ、トークンのカウントがインクリメントされる。現時点におけるトークンカウントが１０２６における最大トークン値ｎよりも小さい場合には、トークンカウントは１０２８においてインクリメントされる。トークンが失われることがあるので、他方端における遷移は、トークンパッシ

ング方式によって許容される許容可能な自由度の範囲内であり、受取られたトークンによって、両サイドが同期状態に戻される。

【０１０４】

トークンカウントがN未満でない場合には、トークンカウントはその最大値である。したがって、システムは遷移をする必要がある。これは、１０３０においてトークンを送出することによって行なわれ、その後、システムがダウンする。これは、ＣｏｎｎＰ→０または、１０２２においてＤとして示される。これにより、ダウンルーチン処理動作が開始される。

【０１０５】

ダウンルーチン処理動作は、アップルーチン処理動作と同様である。タイム・アウトイベントが１０３０で検出される。これは、システムが既にダウンしているので何ら影響を及ぼすことはない。タイム・インイベントが１０３２で検出される。このタイム・インイベントは、もしその遷移を示すよう送ることのできるトークンがあれば、システムをアップ状態に戻すことができる。このルーチンは、ステップ１０４０においてトークンを調べる。利用可能なトークンがなければ、遷移は起こらず、フローは１０２２に戻る。手渡せるトークンがあれば、フローは１０４２に進み、トークンカウントがデクリメントされる。ＣｏｎｎＰの変数は、そのアップ状態に戻り、トークン処理ルーチンが始まる。

【０１０６】

経路Ｘを介したノードＡからノードＢへの各システムは、ＮＥＴＭプロトコルによってこのように特徴付けられる。

【０１０７】

アプリケーションは、ＲＵＤＰの上で実行される。たとえば、プロセスＩＤを有するアプリケーションは、最初に、システムに対して自身の身元を明かす。たとえば、このアプリケーションは、プロセス６として自身を識別しかつプロセス４に対して送信の希望を示す、メッセージを送信することができる。このアイデンティフィケーションは、上述のＣｉ，ｊ，ｋタプルを使用する。ＮＥＴＭは、このオペレーションのための通信経路を決定する。

【０１０８】

実際の通信は、一旦決定されると、いわゆるスライディングウィンドウプロトコルを使用して行なわれる。スライディングウィンドウはよく知られており、たとえば、米国特許番号第５，３０７，３５１号に記載されている。スライディングウィンドウは、データパケットを適切にパッケージングすることによって、高信頼メッセージ通信方式を監督する。スライディングウィンドウは、実質的に、シーケンス番号および確認を管理する。データは高信頼パケットとして送出され、２つ以上のウィンドウが送出される前に受信側が受信を確認することが要求される。受信が正しく確認されると、情報のウィンドウは次の、まだ確認されていない情報のパケットに「スライド」する。

【０１０９】

ＲＵＤＰは、スライディングウィンドウモジュールを使用して実際の通信を行なう。ＲＵＤＰはまた、有効な情報経路を提供するよう、ＮＥＴＭをコールする。ノード間に使用できる経路が２つ以上ある場合には、ＲＵＤＰはそれら使用可能な経路を巡回する。

【０１１０】

ＲＵＤＰはまた、ＮＥＴＭによって提供される情報を使用してシステムを再設定することによって、論理ネットワークとしても機能する。

【０１１１】

基本的なＲＵＤＰのフローチャートを図１１に示す。オペレーションは、ステップ１１００において受信イベントを判断することから開始する。ステップ１１００において受信イベントが受信されなければ、ステップ１１０２において、送信イベントがあったかどうかが判断される。もし送信イベントがなければ、ＬＮＥＴは何もすることがなく、フローは戻って、イベントのチェックを続ける。

【０１１２】

ステップ１１００において受信イベントが検出されれば、フローはステップ１１１０に進む。ステップ１１１０は、データが何らかのＣｉ，ｊ，ｋタプルを示しているかどうかを判断する。もしそうでなければ、ステップ１１１２において誤りが判断される。

【０１１３】

正しいデータが得られれば、そのデータはステップ１１１４において受信されて、ステップ１１１６においてシステムに戻される。

【０１１４】

送信イベントは、送出されるデータおよびイベントを受信するリモートマシンを示す、Ｃｉ，ｊ，ｋの引数を要求する。このためには、リモートマシンに対して動作の引数の１つとして示されるアップチャネルＣｉ，ｊ，ｋが存在するかどうかを、ステップ１１２０で判断せねばならない。もし存在しなければ、ステップ１１２２において接続損失誤りが宣言されるが、より一般的なケースとして、少なくとも１つのアップチャネルが存在すれば、そのアドレスはＣｉ，ｊ，ｋタプルの引数を使用している。その後、プロセスは１１３０で戻る。

【０１１５】

プロセス１１２０は、ＮＥＴＭを使用して、ローカルマシンからリモートマシンへの既存の経路を調べる。したがって、ＮＥＴＭは、ＬＮＥＴがデータ構造を使用する間も、データ構造を維持する。

【０１１６】

情報サーバ

ここに記載するシステムは、情報サーバ、すなわち、リクエストに応じてユーザに情報を提供するサーバ、に特別に応用される。情報サーバは、インターネット（ウェブ）サーバ、ビデオサーバ、または、情報を提供する他のどのような種類のデバイスであってもよい。

【０１１７】

システムは、どのノードも、他のいずれかのノードまたはノードの組合せからそこに記憶されているどのような情報もリクエストすることができる、という意味で、サーバとして使用される。たとえば、２５個の異なるノードから情報を要求するリクエストを行なうこともできる。このシステムは、最も近くにある２５個のノード、または、最も使用頻度の低い２５個のノードを選択することができる。これにより、システムは、過負荷のノードを、それらがあたかも障害を起こしているかのように無視することができる。

【０１１８】

システムがビデオサーバとして使用される場合、送出されるべきビデオの記憶場所としては、システム上のあらゆる場所が可能である。本方式に従えば、ビデオは分散情報として、ネットワークの種々のノード間で記憶されており、特定のネットワークが故障した場合にもビデオ情報が取出せるようにしている。

【０１１９】

サーバシステムは、ビデオを記憶しているノードから、そのビデオが提供されるようリクエストする。このシステムの特別な技術は、特定数の故障によってシステム全体の動作が決して中断されることのないようにする。たとえば、２つのノードが故障しても、記憶情報を得ることはできる。というのも、情報はネットワーク内の他のローケーションに冗長に記憶されているためである。

【０１２０】

別の応用例としてウェブサーバがある。ウェブサーバは、ＴＣＰ／ＩＰプロトコルおよびパケット通信を使用して、インターネットの情報を取得する。やはり、この情報も、分散型サーバ内のどこに記憶されていてもよい。通信または記憶を問わず、どのような２つの障害も、情報の取得を妨げることはない。

【０１２１】

このシステムは、他にも、拡張および修復に応用される。どのノードもどの時点においても取除くことができ、システムの残りの部分はそれでも中断なく動作し続ける。そのノードは、空白のノードに置換されてもよく、その場合、ネットワークは、それが空白の列であると認める列に、冗長データを利用して情報を書込み始める。

【０１２２】

以上に、いくつかの実施例のみを詳細に開示したが、当業者には、開示された実施例の範囲内に他の実施例が存在すること、また、開示した実施例から、本発明を実施するための技術が他にも想定され得ること、が理解されるであろう。

【図面の簡単な説明】

【図１】　最も簡単なネットワークの例を示す、基本的なブロック図である。

【図２】　より多くのスイッチおよびより多くの計算ノードを有する、より

複雑な例を示す。

【図３】　より信頼性の高いネットワーキングの例を示す。

【図４】　フォールトトレラントシステムを示す。

【図５】　このシステムを使用してビデオを記憶する方法の例を示す。

【図６】　システムがどのようにリンク障害を許容することができるかを示す。

【図７】　信頼性の高い通信のための、ソフトウェアアーキテクチャのブロック図である。

【図８】　ネットワーク監視プロセスのための基本的なソフトウェアフローチャートを示す。

【図９】　ネットワーク監視プロセスのための接続プロトコルステートマシンを示す。

【図１０Ａ】　接続のためのデータ構造のフォーメーションを示す。

【図１０Ｂ】　リンクステータスオペレーションのフローチャートを示す。

【図１１】　ＲＵＤＰプロセスのフローチャートを示す。

【図１２】　計算ノードおよびスイッチング要素の可能な配列を示す。

【図１３】　計算ノードおよびスイッチング要素のより高度な配列を示す。

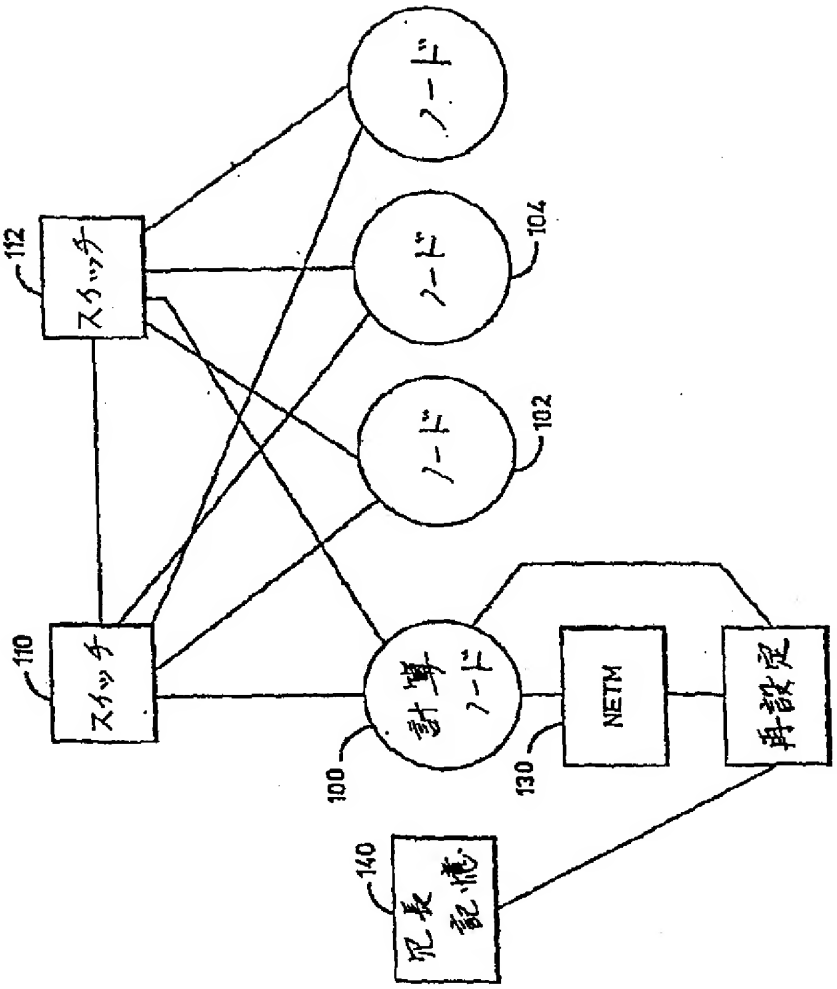
【図１４】　Ｘ符号における、５×５のアレイ符号のための、パリティ行の計算を示す。

【図１５】　Ｘ符号システムの基本的なレイアウトを示す。

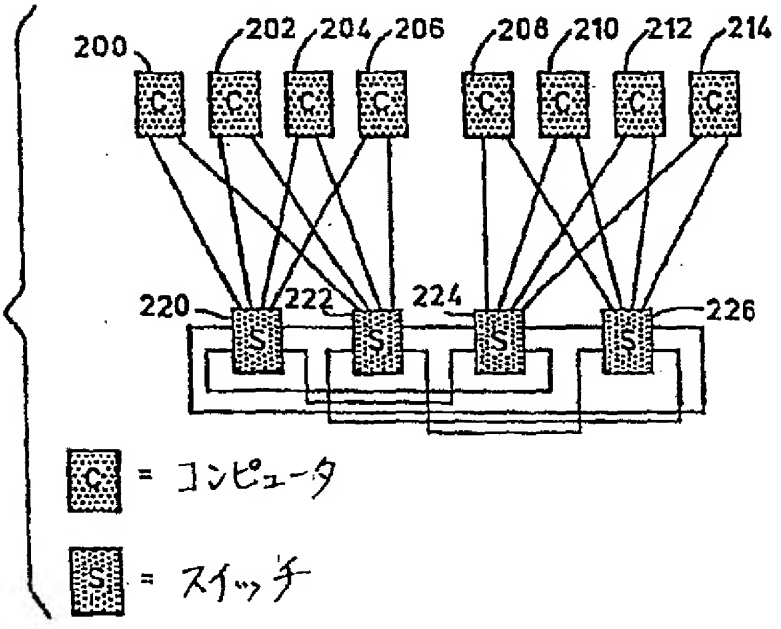
【符号の説明】

１００，１０２，１０４，１０６　計算ノード、１１０，１１２　スイッチ、１３０　ネットワーク監視（ＮＥＴＭ）、１４０　冗長記憶。

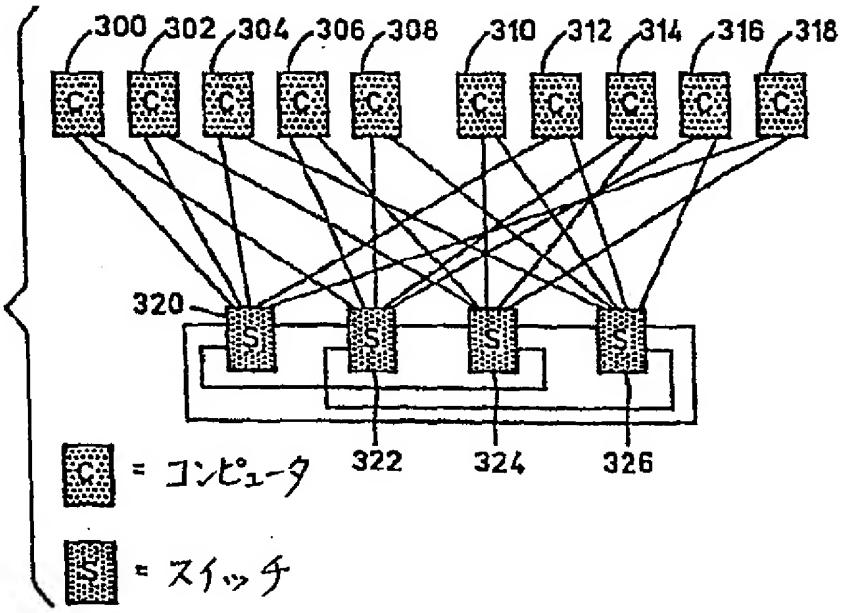
【図１】


ノード
ノード
104
ノード
102
スイッチ
112
スイッチ
110
計算ノード
100
NETM
130
再設定
140
冗長記憶

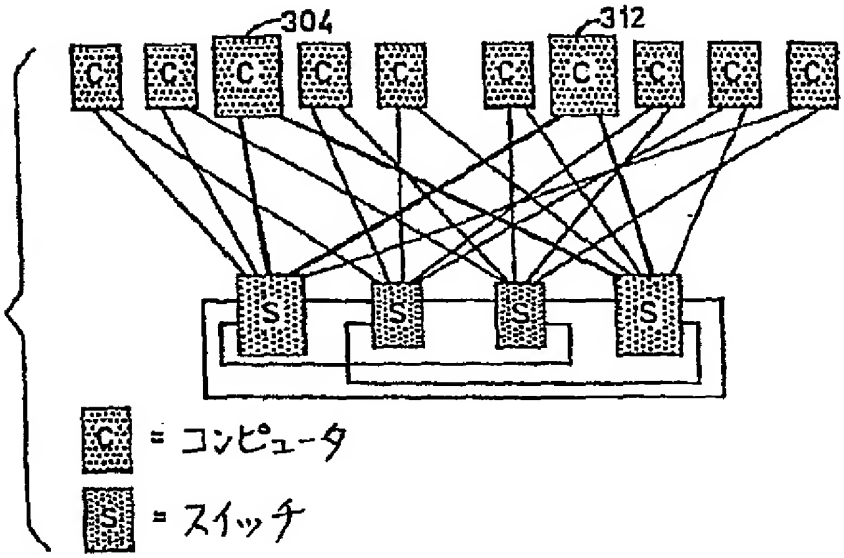
【図２】


200
202
204
206
208
210
212
214
220
222
224
226
C = コンピュータ
S = スイッチ

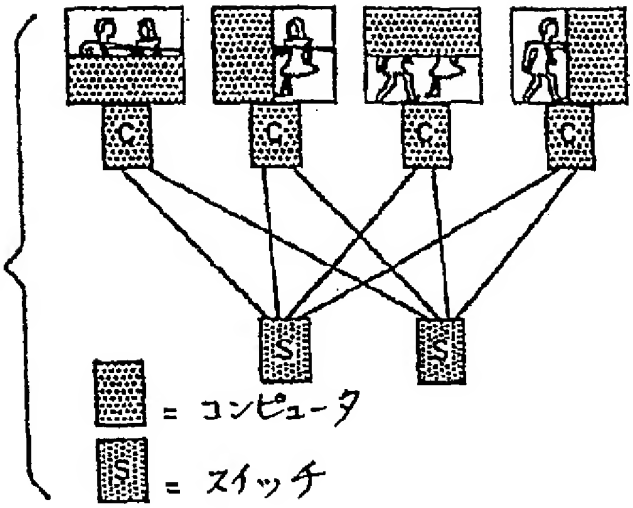
【図３】


300
302
304
306
308
310
312
314
316
318
320
322
324
326
C = コンピュータ
S = スイッチ

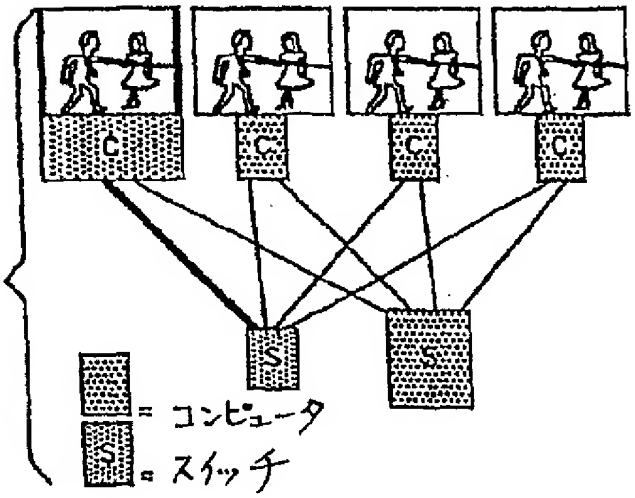
【図４】


304
312
C = コンピュータ
S = スイッチ

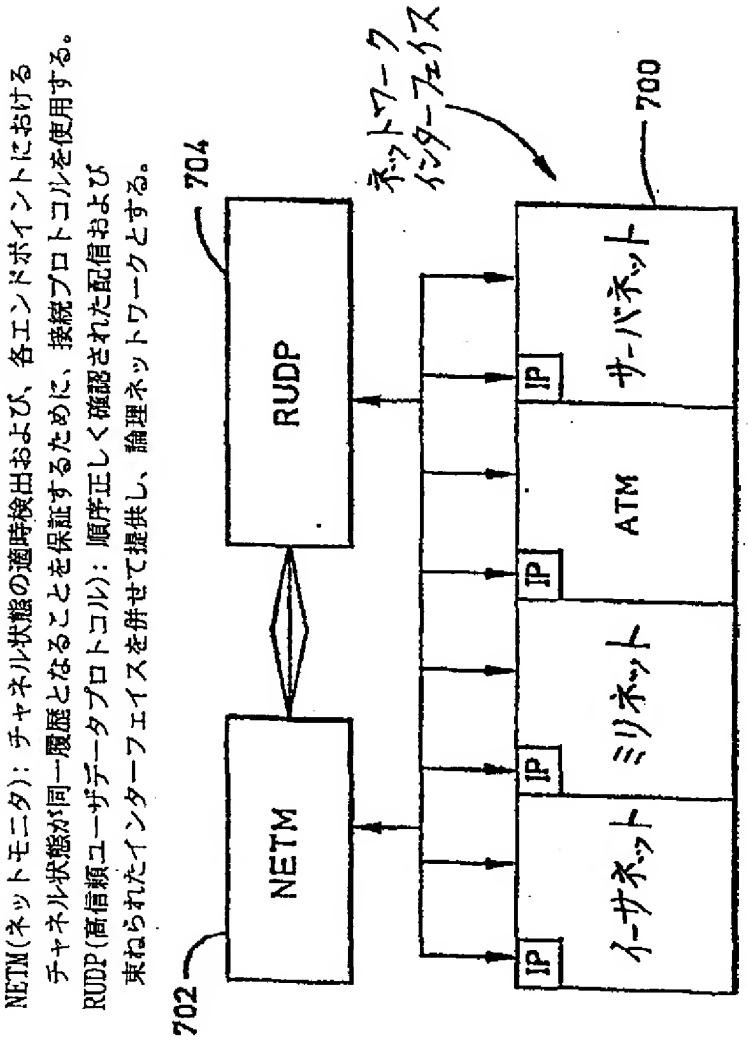
【図５】



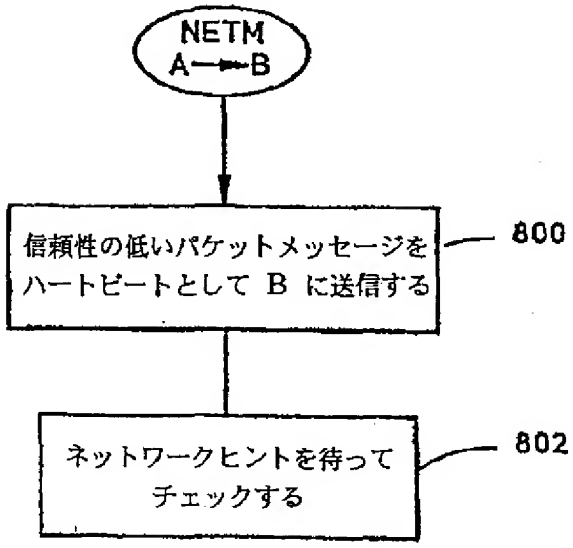
【図６】



【図７】

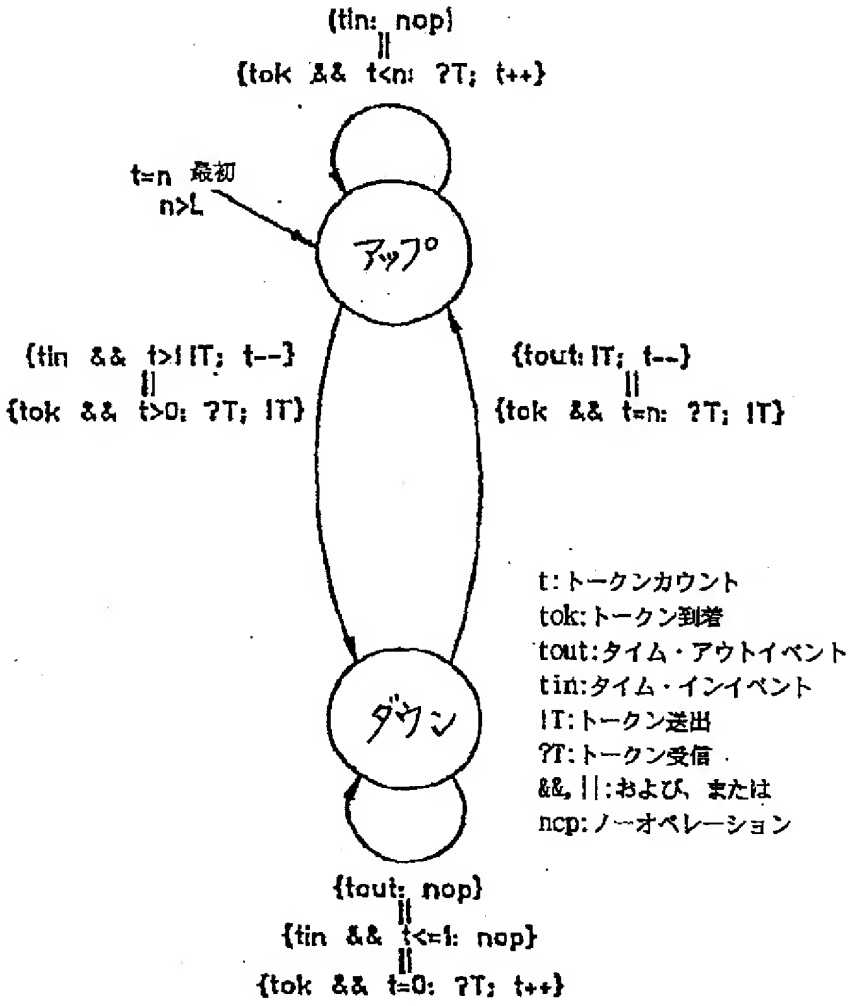

高信頼通信のためのソフトウェアアーキテクチャ
NETM(ネットモニタ)：チャネル状態の適時検出および、各エンドポイントにおけるチャネル状態が同一履歴となることを保証するために、接続プロトコルを使用する。
RUDP(高信頼ユーザデータプロトコル)：順序正しく確認された配信および束ねられたインターフェイスを併せて提供し、論理ネットワークとする。
704
RUDP
NETM
702
ネットワークインターフェイス
700
IP
IP
IP
IP
サーバネット
ATM
ミリネット
ルーザネット

【図８】

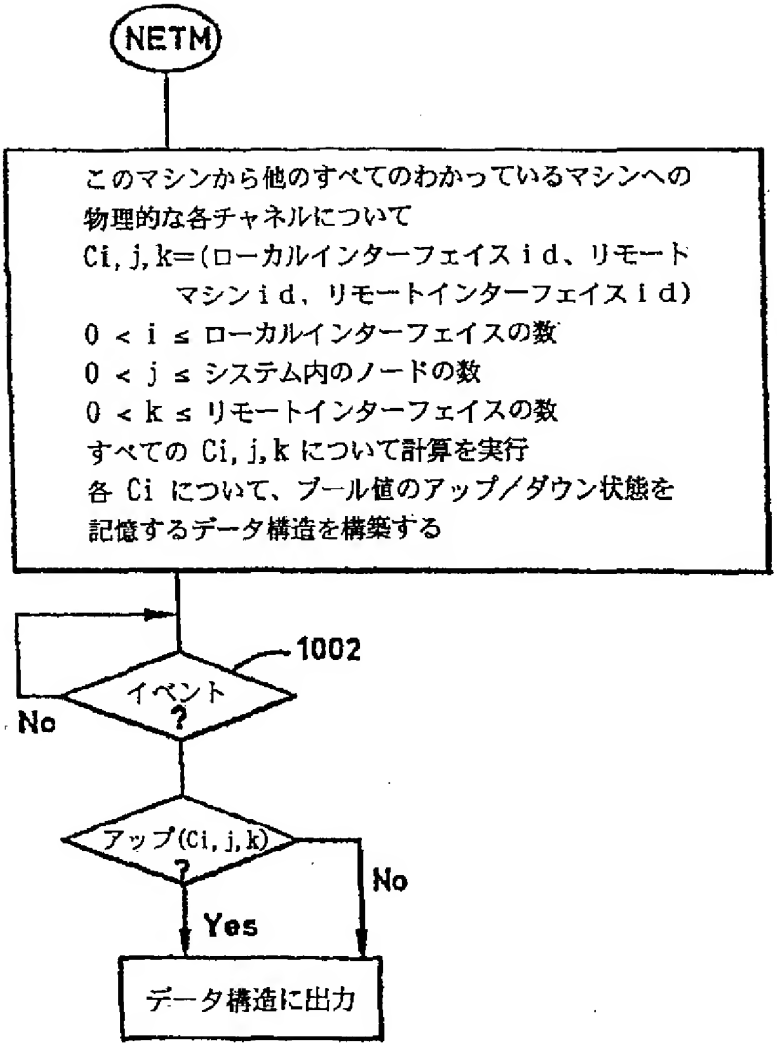

NETM
A→B
信頼性の低いパケットメッセージをハートビートとして B に送信する
800
ネットワークヒントを待ってチェックする
802

【図９】

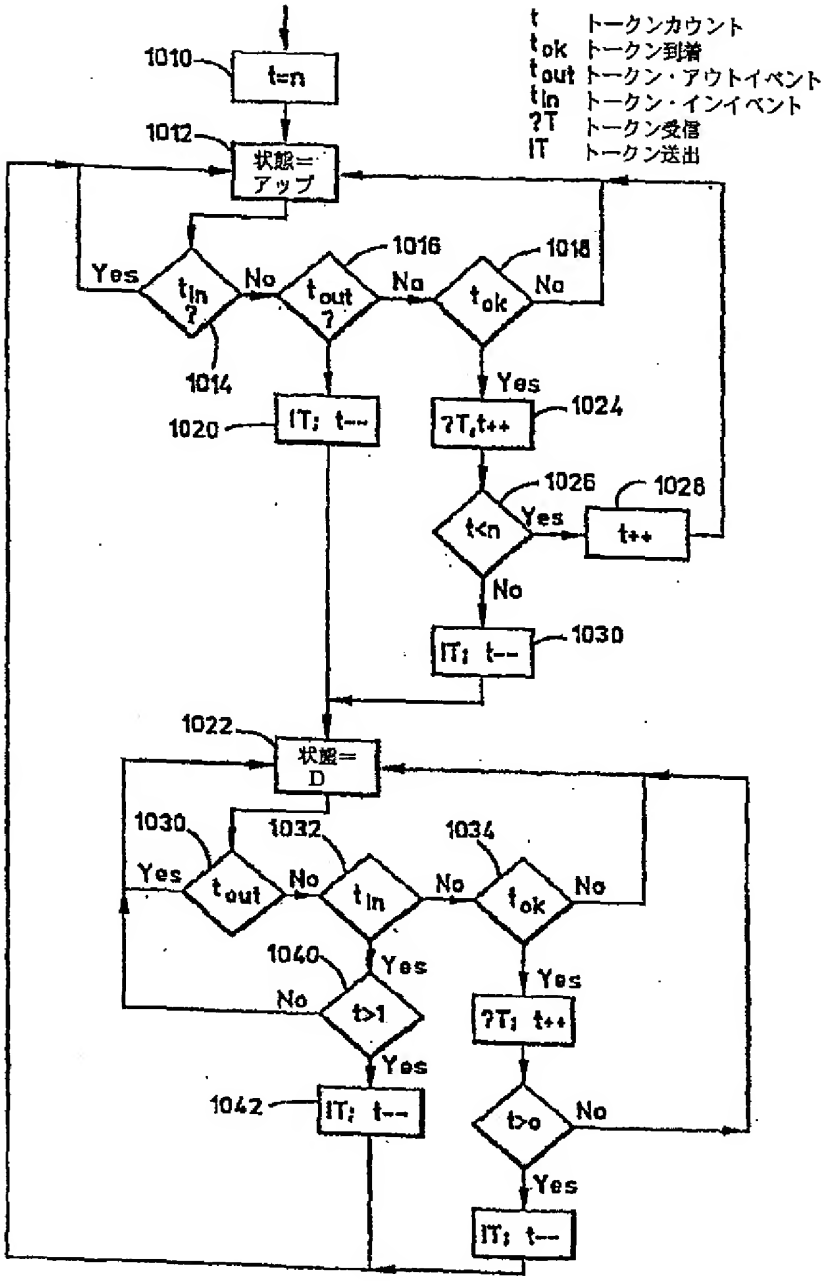
接続プロトコルステートマシン



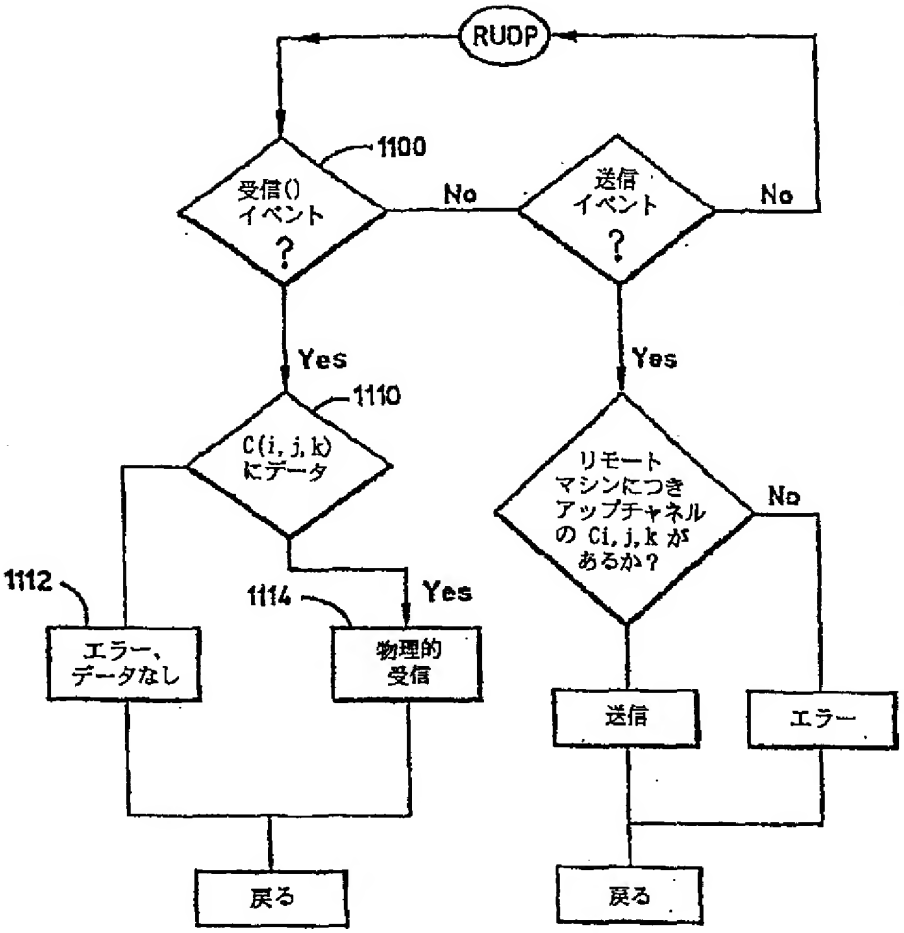
【図１０Ａ】


NETM
このマシンから他のすべてのわかっているマシンへの
物理的な各チャネルについて
Ci,j,k=(ローカルインターフェイスｉｄ、リモート
マシンｉｄ、リモートインターフェイスｉｄ)
0 < i ≦ ローカルインターフェイスの数
0 < j ≦ システム内のノードの数
0 < k ≦ リモートインターフェイスの数
すべての Ci,j,k について計算を実行
各 Ci について、ブール値のアップ／ダウン状態を
記憶するデータ構造を構築する
1002
イベント
?
No
アップ(Ci,j,k)
?
No
Yes
データ構造に出力

【図１０Ｂ】


t トークンカウント
tok トークン到着
tout トークン・アウトイベント
tin トークン・インイベント
?T トークン受信
IT トークン送出
1010
t=n
1012
状態=アップ
1014
1016
1018
Yes
tin ?
No
tout ?
No
tok
No
Yes
1020
IT; t--
?T; t++
1024
1026
t<n
Yes
1028
t++
No
1030
IT; t--
1022
状態=D
1030
1032
1034
Yes
tout
No
tin
No
tok
No
Yes
1040
t>1
No
Yes
1042
IT; t--
?T; t++
t>0
No
Yes
IT; t--

【図１１】



【図１２】

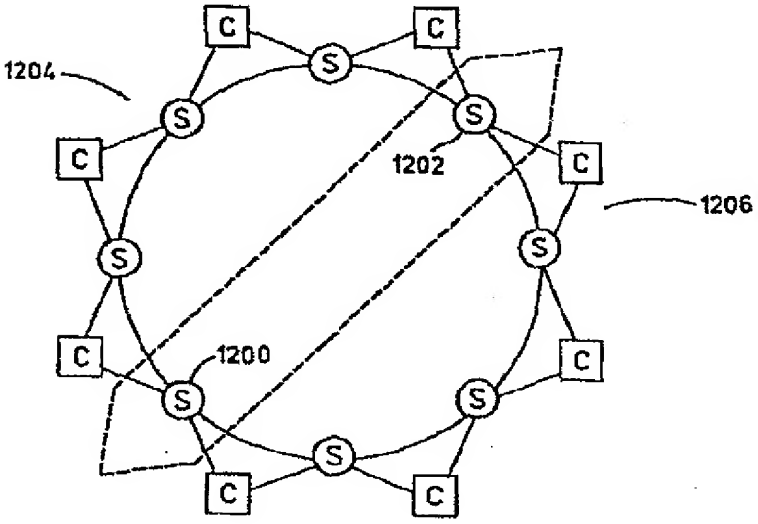
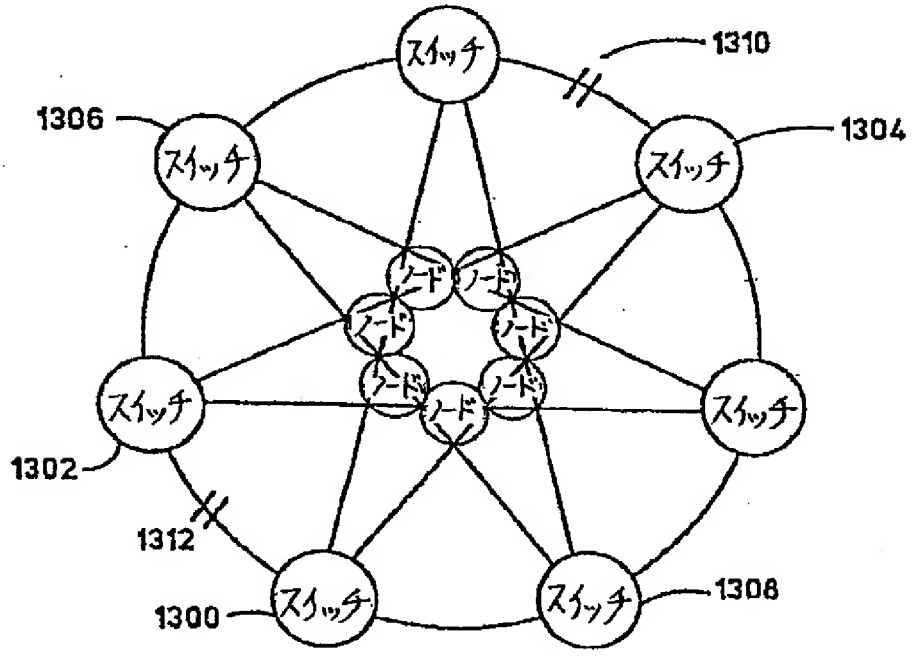


FIG. 12

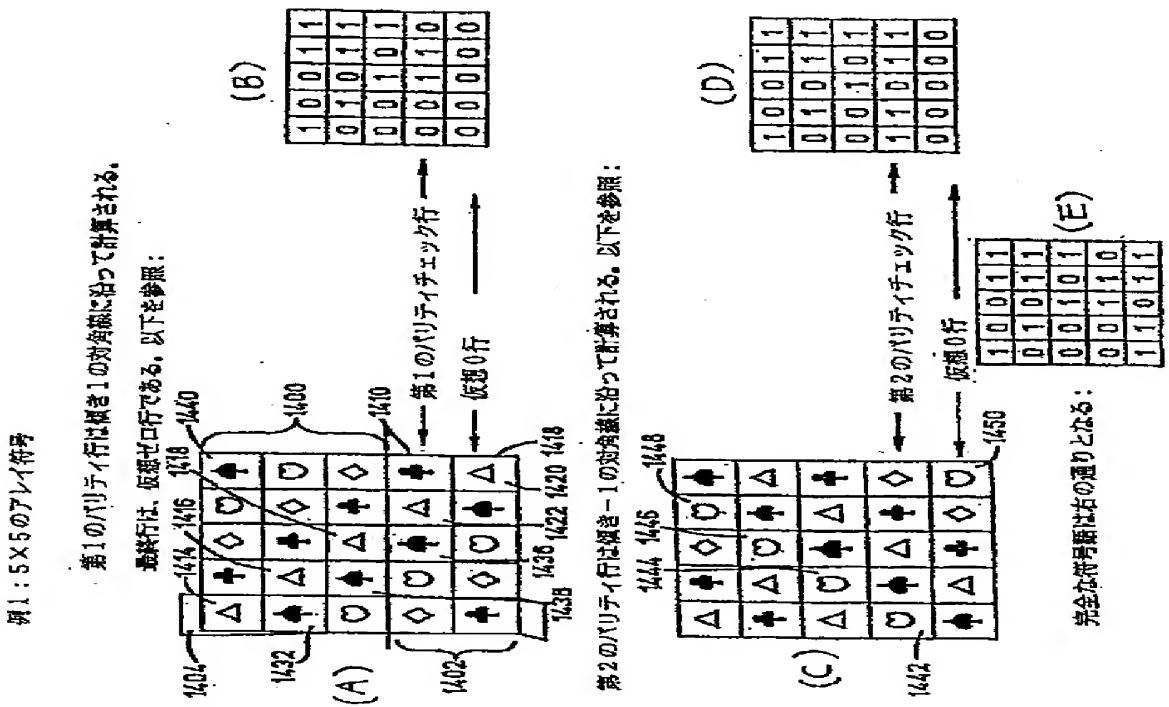
【図１３】

一般的な問題
設定：スイッチおよびノードのネットワーク． 目標：ノード間の通信． 障害：スイッチ、ノードまたはリンクの故障．

具体的な問題
設定：スイッチはパケットを前方に送り、ノードは送らない． 目標：分離されるノードの数を一定に保つ． 障害：スイッチの故障．



【図１４】



【図１５】

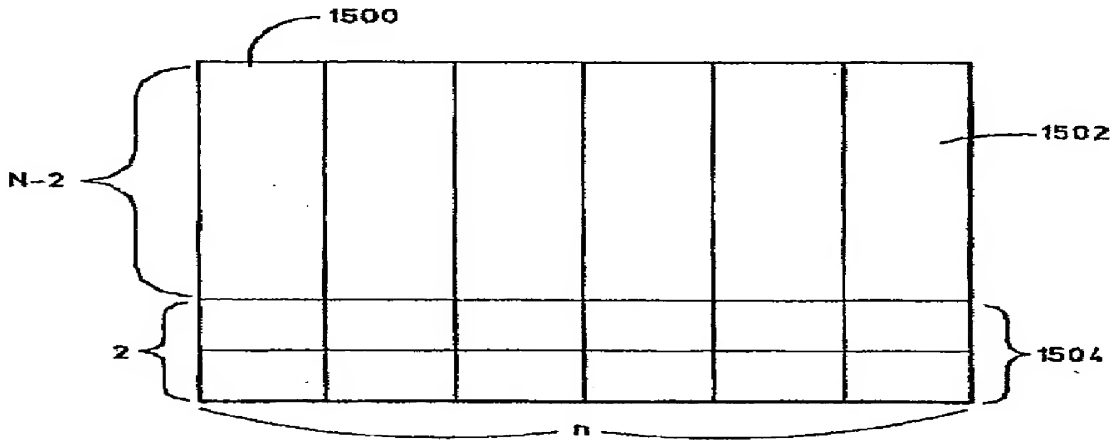


FIG. 15

【手続補正書】特許協力条約第３４条補正の翻訳文提出書

【提出日】平成１２年３月３１日（２０００．３．３１）

【手続補正１】

【補正対象書類名】明細書

【補正対象項目名】特許請求の範囲

【補正方法】変更

【補正内容】

【特許請求の範囲】

【請求項１】　冗長分散型ネットワークシステムであって、

複数のシステムノード（１００，１０２，１０４，１０６）を含み、前記システムノードの各々は少なくとも２つの通信装置および記憶装置を有し、前記記憶装置はいずれか１つのノードにおける記憶情報が他のいずれかのノードにおけるデータと組合されたときに記憶されたデータの再構築を可能にする、ネットワークのための情報の冗長記憶を有し、さらに

前記システムノードの前記通信装置に、いずれか１つのシステムノード内の前記通信装置の各々が複数のスイッチング装置の異なる１つに接続されて、前記システムノードの各々が少なくとも２つの異なる経路のうちの１つを介して互いに通信することができ、したがって冗長通信を提供するように接続される、複数のスイッチング装置（１１０，１１２）と、

ネットワークシステムの動作を妨害し得るシステム機能状態を検出する、各ノード（１００，１０２，１０４，１０６）における検出ルーチンと、

必要に応じてネットワークの冗長性を利用してシステム機能状態を補償するよう、前記通信冗長性または前記記憶冗長性のうち少なくとも１つを使用してネットワークを再設定する、論理ネットワークプロセス（１４０）とを含む、システム。

【請求項２】　前記検出ルーチンは、前記システムノードの少なくとも複数上の機能状態を検出し、前記システムノードの各々上では各前記検出ルーチンによって同じプロトコルが実行される、請求項１に記載のシステム。

【請求項３】　前記システム機能状態は、ネットワーク通信内の障害、メモ

リ記憶内の障害、または、望ましくない結果をもたらす他の種類の障害を含む、請求項１に記載のシステム。

【請求項４】　前記システム機能状態は、ネットワーク通信における障害を含み、前記論理ネットワークプロセスは、接続が異なる接続に変更されるよう指令する、請求項３に記載のシステム。

【請求項５】　前記システム機能状態がメモリ記憶における障害を含むときには、前記論理ネットワークプロセスは、所望の情報が前記冗長データ記憶から得られるよう指令する、請求項４に記載のシステム。

【請求項６】　前記検出ルーチン（７０２）は各ノード（１００，１０２，１０４，１０６）で動作して、ネットワーク内の他のノードに対する接続状態を見る、請求項１に記載のシステム。

【請求項７】　前記検出ルーチン（７０２）は、ネットワークの状態に関するヒントを使用して前記接続状態を判断するよう動作する、請求項６に記載のシステム。

【請求項８】　前記ヒントは、所定の間隔で各前記ノードによって生成されるハートビート信号を含み、前記検出ルーチンは、前記ハートビート信号を受信して前記ヒントの１つとしての前記ハートビート信号の存在または不在を検出するよう動作する、請求項７に記載のシステム。

【請求項９】　トークンパッシングシステムをさらに含み、各ノード（１００，１０２，１０４，１０６）は監視されるチャネルを介して監視されるノードにおけるイベントを判断し、前記イベントを示すよう前記監視されるチャネルを介して前記監視されるノードにトークンを渡し、前記監視されるノードは、前記イベントに基づいた動作を示すよう前記トークンを戻し、各ノードは、該ノードとは異なるノード上で対応するイベントが起きずとも該ノードにおいて起こすことのできるイベントの数を制限するよう、所定数のトークンのみを有する、請求項７に記載のシステム。

【請求項１０】　各前記ノードがネットワークの同じ履歴を見ることを保証する手段をさらに含む、請求項６に記載のシステム。

【請求項１１】　前記接続は、計算ノードのグループが分離されることのな

いように行なわれる、請求項１に記載のシステム。

【請求項１２】　前記スイッチ（１１０，１１２）は、可能な限り最も非ローカルな方法で前記ノードを接続する、請求項１に記載のシステム。

【請求項１３】　前記スイッチは、互いに最も遠い２つのノード間を接続する、請求項１２に記載のシステム。

【請求項１４】　前記接続は、どの２つのノードの故障によってもノードの１グループをノードの他のグループとの通信から分離することのないように行なわれる、請求項１３に記載のシステム。

【請求項１５】　各ノード（１００，１０２，１０４，１０６）は、少なくとも２つの経路によって各他ノードと接続され、

ネットワークモニタ（７０２）をさらに含み、該ネットワークモニタは、各前記ノードにおいて稼動して、前記ネットワークモニタがその上で稼動するローカルノードから各リモートノードへの各接続経路上のすべての接続を監視する、請求項１に記載のシステム。

【請求項１６】　高信頼ユーザデータプロトコルをさらに含み、該プロトコルは、前記ローカルノード上で稼動し、該ローカルノードから他のノードへの通信リクエストを受信し、かつ、前記ネットワーク監視プロセスから経路を判断する、請求項１５に記載のシステム。

【請求項１７】　ノード間の物理的な接続を、異なるノード接続に変更することを可能にする、論理ネットワーク相互接続を使用した前記通信経路の再設定をさらに含む、請求項１６に記載のシステム。

【請求項１８】　前記ノード間の動作可能な接続を判断するネットワークモニタ（７０２）と、稼働中のノードのための情報を処理する高信頼ユーザプロトコルと、前記動作可能な接続に基づいて通信を再設定する論理ネットワークと、をさらに含む、請求項１に記載のシステム。

【請求項１９】　前記記憶装置は、各ノードの各ディスク上の情報の一部のみを記憶する、請求項１に記載のシステム。

【請求項２０】　各ノードの各ディスクは、他のディスク上の情報の特性を示す情報もまた記憶する、請求項１９に記載のシステム。

【請求項２１】　冗長分散型サーバであって、

分散型計算ノード（１００，１０２，１０４，１０６）のアレイを含み、前記計算ノードの各々は各他ノードとは異なる記憶された情報を有し、前記記憶された情報は、いずれか１つのノード内の記憶された情報が他のいずれか１つのノード内の記憶された情報と組合せられたときに記憶されたデータを再構築することができるように、前記計算ノード間で冗長となっており、さらに

前記計算ノードのアレイに接続され、前記計算ノードのアレイ間に冗長通信経路を提供し、どのような所定数のネットワーク障害もシステムの残りのノードの動作に影響を及ぼすことのないように動作する、スイッチングシステム（１１０，１１２）を含み、

前記計算ノードの各々はネットワークステータスを判断する同じプロトコルを実行し、よって、各前記計算ノードが同じネットワーク履歴を見るようにし、各前記計算ノードは、ネットワークシステムの動作を妨害し得るシステム機能状態を検出する検出ルーチンを含む、サーバ。

【請求項２２】　前記各ノード上に記憶される情報は、所望の情報のすべてではない一部のみを記憶し、どの２つのノードも同じ情報を記憶することはない、請求項２１に記載のサーバ。

【請求項２３】　前記記憶される情報は、情報部分および冗長部分を含み、前記冗長部分は他のノードのための情報部分のみを示す情報である、請求項２２に記載のサーバ。

【請求項２４】　前記冗長部分は、複数の前記ノードが前記情報部分を形成するようアレイへと配列されているアレイ符号から形成され、かつ、前記冗長部分は、前記アレイの対角線方向に沿ったチェックサムで形成される、請求項２３に記載のサーバ。

【請求項２５】　冗長性を提供するようネットワークを操作する方法であって、

集合的にシステムデータを記憶する複数のノード（１００，１０２，１０４，１０６）から分散された読出しを実行する、制御プロセスを実行するステップを含み、各ノードは、生のデータおよび、前記各ノード以外のノード内に記憶され

ている生のデータを示す冗長データを記憶して、いずれか１つのノード内の記憶情報が他のノードにおけるデータと組合せられたときに記憶されたデータを再構築することができるようにし、各ノードは、ネットワークシステムの動作を妨害し得るシステム機能状態を検出する検出ルーチン（７０２）を含み、さらに

分散された読出しを行なうステップを含み、該読出しを行なうステップは、

ノードの可用性に関するパラメータを判断するステップおよび、

前記パラメータが可用性を示す場合には前記複数のノードから前記生のデータを読出し、また、前記パラメータが可用性を示さない場合には前記複数のノードよりも少ない数のノードから前記生のデータおよび前記冗長データを読出すステップを有する、方法。

【請求項２６】　誤り訂正符号を示す生の情報および冗長情報を複数の情報ノードに記憶するステップと、

前記情報ノードのユーザビリティを示すパラメータを判断するステップと、

前記パラメータが前記複数のノードが使用可能であることを示す場合には前記複数のノードから前記生の情報を読出し、かつ、前記パラメータが前記複数のノードのうち少なくとも１部分が使用可能ではないことを示す場合には前記複数のノードよりも少ない数のノードから前記生のデータおよび前記冗長データの両方を読出すステップとをさらに含む、請求項２５に記載の方法。

【請求項２７】　ノードからの情報を表わす、アレイの各列を形成することによって、情報のアレイを形成するステップと、

データを示す生の情報を含む、各列の生の部分を形成するステップと、

冗長性に関する情報を示す、冗長情報を形成するステップとをさらに含み、前記冗長情報は、前記各ノード以外の他のノードに関する情報を示し、該情報は、前記他のノードから情報を得る特定の形状のエンベロープに沿って取られる、請求項２５に記載の方法。

【請求項２８】　前記エンベロープは、前記アレイの縁を超えて他ノードに延びる対角線である、請求項２７に記載の方法。

【請求項２９】　アレイの列に各ノードをマッピングするステップと、

前記アレイの列から冗長情報の２つの行を形成し、前記２つの行を前記列内に

位置付けることによって、N－2 ×N個の情報シンボルおよび２×N個の冗長情報シンボルを含む、結果として得られるN×Nのアレイを形成するステップとを含み、前記パリティシンボルは以下の式に従って構築される、請求項２５に記載の方法。

【数１】

$$C_{n-2,i} = \sum_{K=0}^{n-3} C_{k,(i+k+2)_n}$$

$$C_{n-1,i} = \sum_{K=0}^{n-3} C_{k,(i-k-2)_n}$$

式中、 $i = 0, 1, \cdots, n-1$ および $\langle x \rangle_n = X \bmod n$ である。

【請求項３０】 該記憶装置はビデオ情報を記憶する、請求項１に記載のネットワーク。

【手続補正２】

【補正対象書類名】明細書

【補正対象項目名】０００８

【補正方法】変更

【補正内容】

【０００８】

この場合、いずれかの接続またはバスにエラーがあっても、他方のバスを介して正常な動作を続けることができる。２つの冗長バスおよび２つの冗長サーバを有するシステムを、デュアルバス、デュアルサーバと呼ぶ。このようなデュアルバス、デュアルサーバシステムは、単一のネットワーク障害は許容することができる。しかし、このようなシステムでは、通常、各サーバ上にすべての情報の複製を作成しなければならない。

システムの冗長性を提供する試みは、「フレキシブルなプロトコル能力を有する高速スイッチングシステム」と題された公報ＥＰ－Ａ－０３６６９３５号に記載されているシステムを含む。この文献は、複数のスイッチング面を含んで、１つのスイッチング面またはリンクが障害を起こした場合にも、残りのスイッチング面を使用することによって稼動し続ける、高速のスイッチングシステムを開示

している。このスイッチング面は特定の種類の、特定の目的のためのものである。さらに、このＤ１の文献は、システムの再構成およびデータの復元に関するものではない。

【国際調査報告】

INTERNATIONAL SEARCH REPORT

Inter nal Application No
PCT/US 98/20532

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 G06F15/173 H04N7/173 G06F11/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 G06F H04N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	M.M. BUDDHIKOT ET AL.: "Design of a large scale multimedia storage server" COMPUTER NETWORKS AND ISDN SYSTEMS, vol. 27, no. 3, December 1994, pages 503-517, XP002093312 The Netherlands see page 508, right-hand column, line 1 - line 16 —/—	1,21

[X] Further documents are listed in the continuation of box C. [X] Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
12 February 1999	26/02/1999

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax (+31-70) 340-3016	Absalom, R

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Inter nal Application No.
PCT/US 98/20532

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
A	N.J. BODEN: "Myrinet: A Gigabit-per-Second" IEE MICRO, vol. 15, no. 1, February 1995, pages 29-36, XP000501486 los alamitos, ca, usa cited in the application see page 31, right-hand column, line 12 - line 16; figure 2 see page 34, right-hand column, line 4 - line 17	1
Y	US 5 579 475 A (BLAUM ET AL.) 26 November 1996 cited in the application	28,29
A	see the whole document	
X	P.C. WONG ET AL.: "Redundant Array of Inexpensive Servers (RAIS) for On-demand Multimedia Services" IEEE INTERNATIONAL CONFERENCE ON COMMUNICATIONS, vol. 2, 8 June 1997, pages 787-792, XP000742048 Montreal, Canada	27
Y	see the whole document	28,29
A	US 5 612 897 A (REGE) 18 March 1997 see the whole document	1-30
X	US 5 630 007 A (KOBAYASHI ET AL.) 13 May 1997 see the whole document	25,26
A	WO 91 14229 A (SF2 CORPORATION) 19 September 1991 see page 1 - page 6, line 24; figure 1	1,30
A	S. NAKAMURA ET AL.: "Distributed RAID style video server" IEICE TRANSACTIONS ON COMMUNICATIONS, vol. e79-b, no. 8, 1908 - August 1996, pages 1030-1038, XP000628640 japan	
A	A. COHEN ET AL.: "Segmented Information Dispersal (SID) for fault-tolerant video servers" PROC. OF THE SPIE, vol. 2604, 23 October 1995, pages 58-69, XP000578588 usa	

-/--

1

Form PCT/ISA/210 (continuation of second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Inter nal Application No
PCT/US 98/20532

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	R. TEWARI: "High Availability in Clustered Multimedia Servers" PROC. 12TH CONFERENCE ON DATA ENGINEERING, 26 February 1996, pages 645-654, XP000632617 New Orleans, usa	

Form PCT/ISA/210 (continuation of second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Information on patent family members

Inter nal Application No
PCT/US 98/20532

Patent document cited in search report		Publication date	Patent family member(s)		Publication date
US 5579475	A	26-11-1996	US	5271012 A	14-12-1993
			DE	69408498 D	19-03-1998
			EP	0632376 A	04-01-1995
			JP	2750316 B	13-05-1998
			JP	7028710 A	31-01-1995
			EP	0499365 A	19-08-1992
			JP	2514289 B	10-07-1996
			JP	4310137 A	02-11-1992
			EP	0519669 A	23-12-1992
			US	5351246 A	27-09-1994
US 5612897	A	18-03-1997	NONE		
US 5630007	A	13-05-1997	JP	8329021 A	13-12-1996
			GB	2299424 A,B	02-10-1996
WO 9114229	A	19-09-1991	US	5388243 A	07-02-1995
			AU	7486091 A	10-10-1991
			CA	2077447 A	10-09-1991
			EP	0518965 A	23-12-1992

フロントページの続き

(51)Int.Cl.[7]	識別記号	FI		テーマコード(参考)
H04L 12/28		H04N 7/173	620D	
H04N 7/173	620	H04L 11/20	C	

(81)指定国 EP(AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OA(BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG), AP(GH, GM, KE, LS, MW, SD, SZ, UG, ZW), EA(AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, UZ, VN, YU, ZW

(72)発明者 ボーシアン，バスケン
カナダ、エイチ・４・ジェイ　１・エックス・４　ケベック州、モントリオール、11675、ラビン・アパートメント・６

(72)発明者 ファン，チェンゴン
アメリカ合衆国、91106　カリフォルニア州、パサデナ、エス・メリディス・アベル、156、アパートメント・115

(72)発明者 レマイエウ，ポール
アメリカ合衆国、91106　カリフォルニア州、パサデナ、イー・デル・マー・ブールバード、1032、アパートメント・301

(72)発明者 リーデル，マーカス・デイビッド・ダニエル
カナダ、ジェイ・４・エックス　１・ティ・２　ケベック州、ブロッサード、ストラウス、1015

(72)発明者 シュ，リハオ
アメリカ合衆国、91106　カリフォルニア州、パサデナ、サウス・ウィルソン・アベニュ、307、アパートメント・５